Fall 12-14-2018

Development of Preclinical Magnetic Resonance Imaging Relational Database and Interactive Analytical Tool for Diffusion Tensor Imaging

Ahmad Tanwir
University of Nebraska Medical Center

Development of Preclinical Magnetic Resonance Imaging Relational Database and Interactive Analytical Tool for Diffusion Tensor Imaging

by

Ahmad Tanwir

A THESIS

Presented to the Faculty of

the University of Nebraska Graduate College

in Partial Fulfillment of the Requirements

for the Degree of Master of Science

Biomedical Informatics

Graduate Program

Under the Supervision of Prof. Chittibabu Guda
and
Dr. Balasrinivasa R. Sajja

University of Nebraska Medical Center

Omaha, Nebraska

November, 2018

Supervisory Committee:

Dr. Peter Wolcott

Dr. Mahbubul A Majumder

ACKNOWLEDGEMENT

First of all, I am grateful to my God for the good health and wellbeing that were necessary to complete this long journey. I am especially indebted to my supervisors, Prof. Chittibabu Guda and Dr. Balasrinivasa R. Sajja, for all the selfless time, wisdom, care, and unparalleled support that kept me going. Their unwavering guidance helped me in all the time of research and writing of this thesis.

I whole-heartedly express my sincere thanks to all the members of my supervisory committee, Dr. Peter Wolcott and Dr. Mahbubul A Majumder, not only for their insightful comments and encouragement, but also for the critical question, which incented me to widen my research from various perspectives.

I would like express gratitude to all the members of Bio-imaging core facility. I specially thank Dr. Yutong Liu, Melissa Mellon, Dr. Mariano Uberti, and Dr. Lirong Xu for their helps and supports. Besides that I would like to remember my first supervisor late Prof. Michael D. Boska for his contribution to change my philosophy about the life and science.

I also thank my parents, late Prof. Ruhul Amin and Kohinoor Begum and my in-laws for the unceasing encouragement, support and endless prayer from 8000 miles away from USA.

Last, but not the least, I am grateful and dedicate this thesis to my life-partner, Rahat Jahan who supported me each and every moment and being patient more than me throughout this entire journey. I can clearly shout that, without her tremendous devotion and sacrifices, the journey would not be possible at all. Lastly, I thank my 2.9 years old son, Umair Ahmad for being the most lovable and understanding boy ever in the universe.

Development of Preclinical Magnetic Resonance Imaging Relational Database and Interactive Analytical Tool for Diffusion Tensor Imaging

Ahmad Tanwir, M.S.

University of Nebraska Medical Center, 2018

Co-Advisors: Chittibabu Guda, Ph.D. and Balasrinivasa R. Sajja, Ph.D.

Databases have become an essential component for any organization to collect and manage the information. Similar to other domains, by implementing a database management system (DBMS), data storage, access, analysis and updates can be automated and optimized in the biological, medical and preclinical research areas. This work demonstrated the development of a preclinical database, which contains the processed results from different magnetic resonance imaging (MRI) techniques with relevant biological information. In addition, we designed and implemented an interactive analytical tool to format and analyze diffusion tensor imaging (DTI) data and automatically generate results in the forms of box plot, bar plot, line plot with statistical summary. The performance of the newly built analytical tool was tested to determine the accuracy and robustness by following functional-unit test procedures. Microsoft SQL Server 2016 Express edition was used to develop the database, and an open source programming language *R* was used to develop the interactive analytical tool. Results showed that implementation of DBMS and the interactive analytical tool has drastically reduced data analysis time. While the conventional analytical method required more than three days to format, process, analyze and produce results from one study dataset, the DTI analytical tool returned the same results within 10 minutes. This database and the accompanying tool will be very effective for enhancing the speed of data analysis of diffusion tensor imaging data.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ADC	Apparent Diffusion Coefficient
CSF	Cerebrospinal Fluid
DBMS	Database Management System
DBS	Database System
DTI	Diffusion Tensor Imaging
DWI	Diffusion Weighted Imaging
EHR	Electronic Health Records
EPI	Echo Planar Imaging
ERD	Entity-Relation Diagram
FA	Fractional Anisotropy
IACUC	Institutional Animal Care and Use Committee
MD	Mean Diffusivity
MR	Magnetic Resonance
MRI	Magnetic Resonance Imaging
PI	Principal Investigator
RA	Relative Anisotropy
RDBMS	Relational Database Management System
RF	Radiofrequency
ROI	Region of Interest

SQL	Structured Query Language
TBI	Traumatic Brain Injury
TE	Echo Time
TR	Repetition Time
WM	White Matter

CHAPTER 1: INTRODUCTION

In the era of advanced information technology, “data” is considered as the core of any organization. Data is the group of facts and statistics without any insights that can be in structured and unstructured form. “Information” contains the data and the related facts. Information can be obtained from data by appropriate processing and analyses. Effective use of knowledge, which is a combination of information and insights, has become the primary strategy for any organization or institution. The importance of collecting and storing the data is immense since it is not possible to obtain knowledge without having readily access to proper and complete data. Database System (DBS) is one of the most popular approaches to store and manage the data (Hoffer et al., 2011). A database system consists of one or many databases and a database management system (DBMS), which is responsible for creating the database, inserting the data into the database, updating and deleting the data as needed, and an application software with users (Figure 1.1). A database application is a set computer programs that serves as the interface between the user in the front-end and the DBMS in the back-end (Kroenke & Auer, 2012).

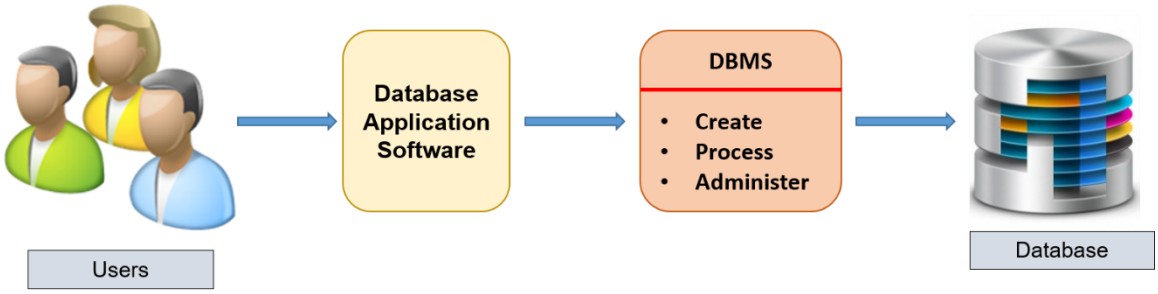


Figure 1.1: Components of the database system (DBS). Users communicate to the database through the application software and DBMS to extract the data (Kroenke & Auer, 2012).

1.1 Database

The database is a collection of logically related data to provide efficient retrieval. The collected data could be in any number of formats like electronic, printed, graphic, audio, statistical, etc. A database doesn't have any limitation on size. It can be very large and complicated. Three fundamental terms of a database are "entity", "attributes" and "relationship". An entity can be defined by a person, a place, an object, an event or a concept that the users want to track. An attribute is a property of an entity. A specific entity will have a value for each of its attributes. Two or more entities are connected to each other by a relationship. There are three types of relationships between entities: 1) one to many (1:M), 2) many to many (M:N), and 3) one to one (1:1). Depending on the connection between the entities and the requirements, the database architecture defines the relationships.

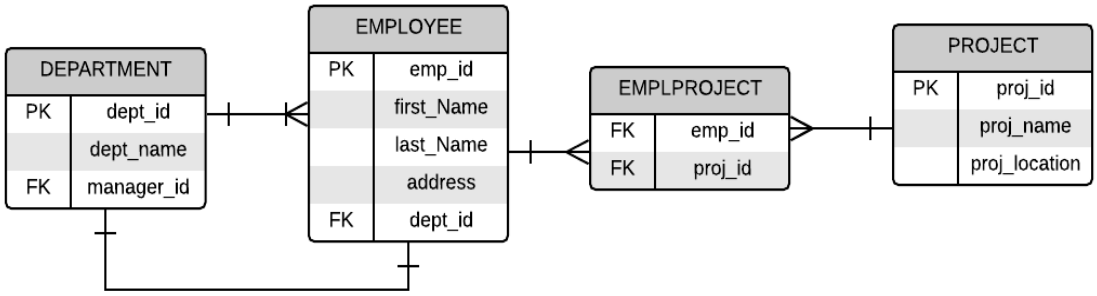


Figure 1.2: Relationship between "DEPARTMENT", "EMPLOYEE", and "PROJECT" entities.

Figure 1.2 shows an example of the relationship between three entities "DEPARTMENT", "EMPLOYEE", and "PROJECT" in an organization. The "DEPARTMENT" entity has three attributes. "dept_id" is the unique value assigned to every department in the organization, "dept_name" is the name of the department, and "manager_id" is the id of the department manager. "emp_id", "first_name". "last_name", "address", and "dept_id" are the attributes

of “EMPLOYEE” entity. The entity “PROJECT” has “proj_id”, “proj_name” and “proj_location” as attributes. A department is related to all employees who work for the organization, and one employee should be assigned to only one department. In this case, there is a one-to-many (1: M) relationship. There is a one-to-one (1:1) relationship between the DEPARTMENT (dept_id) and EMPLOYEE (manager_id) as based on the business rules state that only one manager should manage one department. A many-to-many relationship occurred between EMPLOYEE and PROJECT entity. For example, an employee must work in one or multiple projects and one project must have one or multiple employees. This many-to-many relationship issue is resolved by creating a composite entity which contains “emp_id” and “proj_id”.

1.1.1 Data model

The data model is an illustration of the data elements and the relationships between them. Data modeling is the process of building a simple illustration of a complex system by using text and symbols to represent the data. Usually, it is built during the analysis and design phase of a database to ensure that the requirements are fully understood. Data model structure helps to define the relational tables, attributes, primary and foreign keys, and stored procedures. There are three steps in data modeling. The first one is conceptual modeling, the second one is logical modeling and the last one is physical modeling. Well-documented, and datialed conceptual, logical and physical data models allow stakeholders to identify errors and make changes before starting database development.

1.1.1.1 Conceptual Data Model

Conceptual data model design is the first step in developing a database. The modeler collects all the requirements from the stakeholders and creates the entity relation diagram (ERD) by defining entities, attributes and the associations between entities. In that ERD,

no information is included about the primary key, foreign key, and the relationship between the entities.

1.1.1.2 Logical Data Model

The logical model is a refined version of the conceptual model. In the logical model, information about the attributes, their data types and the relationship among the entities involved based on the business rules, including the cardinality are provided. The advantage of the logical data model is to provide a foundation to form the base for the physical model. However, the modeling structure remains generic. At this data modeling level, primary or secondary key may be defined (Hoffer et al., 2011).

1.1.1.3 Physical Data Model

The physical data model is the final representation of the database design. A physical database model includes all table structures, including column name, column data type, column constraints, primary key, foreign key, and relationships among tables. There are a few steps involved to convert a logical model into a physical model. Those steps are mentioned below:

- Convert entities into tables.
- Convert relationships into primary and foreign keys.
- Convert attributes into columns.
- Modify the physical data model based on physical constraints or requirements.

Examples of three data models are shown in figure 1.3. The conceptual model only included the entity, attributes and the symbol of the association between “DEPARTMENT” and “EMPLOYEE”. In the logical model, the data type and the relationship are included. The relationship between “DEPARTMENT” and “EMPLOYEE” is a one-to-many relationship which means one department can have one or many employees, and one

employee must work for only one department. The physical model added the primary key, foreign key constraints and data types of individual columns.

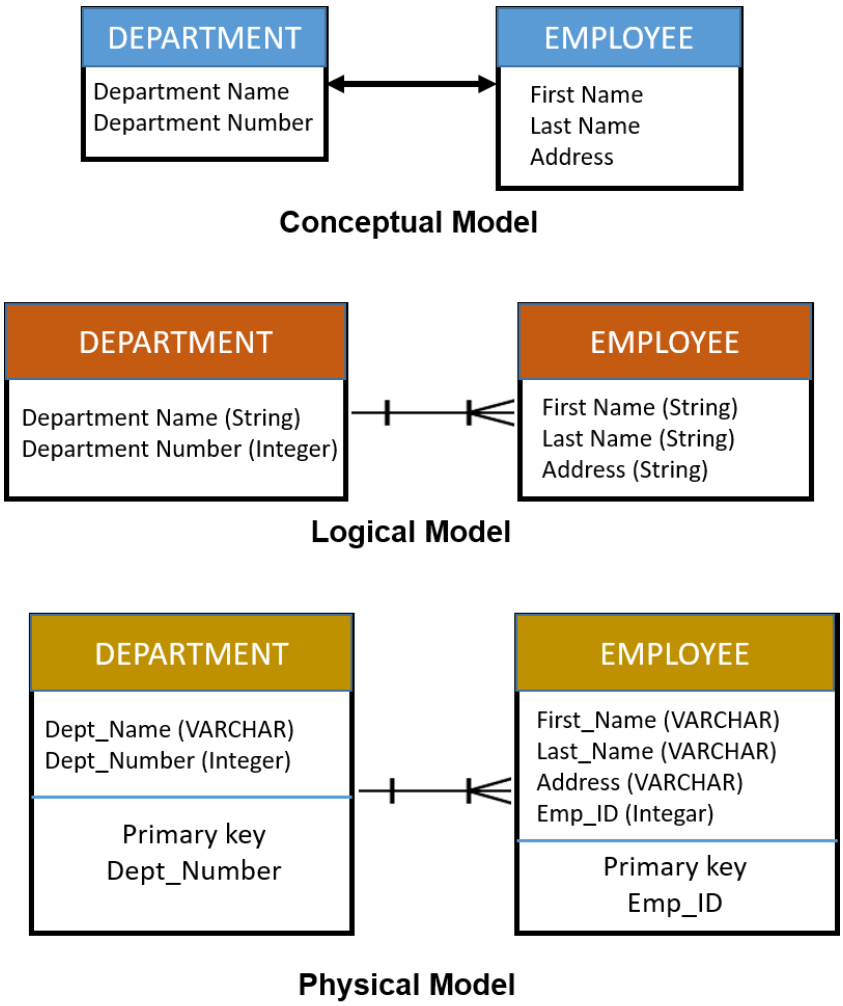


Figure 1.3: Different data models in database development process.

1.1.2 Database models

A database model describes the logical structure of a database, including the relationships and constraints that describe how data can be stored and accessed. Individual database models are designed based on the requirements and concepts of the project. Brief descriptions of a few commonly used models are given below.

1.1.2.1 Relational Model

One of the most commonly used database models is the relational model. The foundation of a relational model is based on the mathematical concept named “relation”, which is comprised of columns and rows as in a matrix. Each row in the relation is named as “tuple”. The relation is also known as a table. Each column of the relation defines an attribute. Those tables are related to each other with a common attribute, which is unique in that table. To implement a relational database, it is required to use a “relational database management system (RDBMS)”. Relational model also defines the relationship between the tables as one-to-one, one-to-many, and many-to-many depends on the requirements and the relationship. Another useful feature of a relational database is the structured query language (SQL). By using the SQL, a user can easily communicate with the database. RDBMS uses SQL to convert the user query into instructions to access the requested data (Coronel et al., 2011).

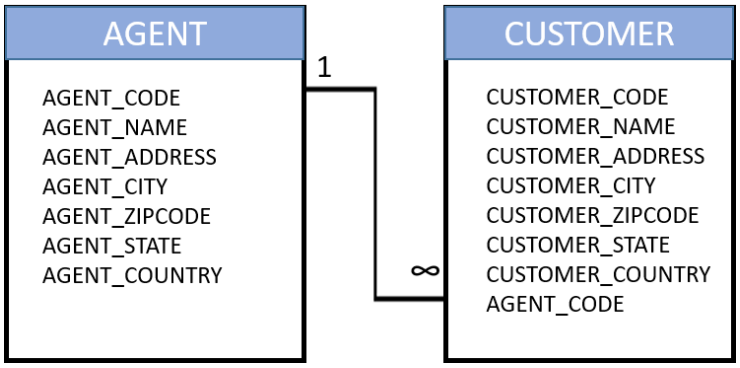


Figure 1.4: A relational diagram.

Figure 1.4 shows an example of a relational data model. Table “AGENT” and “CUSTOMER” are connected through a one-to-many (1: M) relationship by “AGENT_CODE” which is a unique attribute in the table “AGENT”.

1.1.2.2 Entity-Relationship Model

The entity-relationship model represents the entities and their relationships graphically in a database. Entity-relational model diagram (ERD) is the graphical representation of an entity-relationship model, which includes all the components of the database. The primary key and the foreign key are also mentioned in the ERD. There are two types of notations to present the relationship between the two entities. One is Chen notation and another one is Crow's Foot notation (Coronel, C., Morris, S., & Rob, P. (2011).).

An example of an ERD is presented below (Figure 1.5). Table "DEPARTMENT" and table "EMPLOYEE" is related to the key "dept_id" which is the primary key of the table "DEPARTMENT". There is a one-to-many (1: M) relationship between these two tables. The relationship defined that one department can belong to many employees, but one employee must belong to only one department. The figure also showed the example of Chen Notation and Crow Foot Notation.

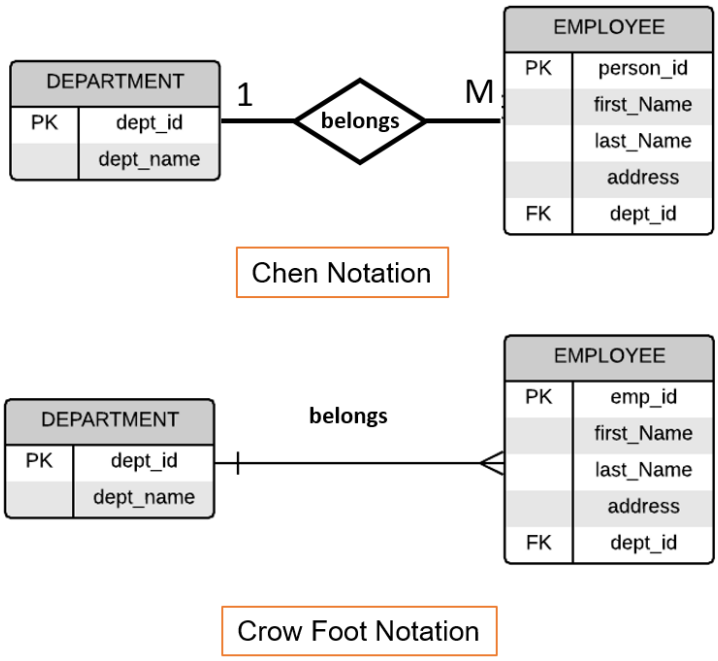


Figure 1.5: Entity- relationship diagram for tables "DEPARTMENT" and "EMPLOYEE", the relationship is shown by using "Chen Notation" and "Crow Foot Notation".

The concept of DBMS is widely used in biological and medical fields. In patient care, physicians collect the data from electronic health records (EHR), medical history, patient health records, patient portals, electronic patient diaries, and wearable fitness tracking devices. Those data have significant influence on the treatment pathway for a patient by providing insight to the physician (Martin 2008). To visualize the mammoth amount of data in a short period, physicians use custom-made visualization tools to observe and monitor patients' activities over time. The visualization tool extracts real-time information from different healthcare databases and provides updates in the form of graphs and alerts (Badgeley et al., 2016).

However, implementation of such DBMS or interactive tools are not so common in preclinical research settings. Preclinical research has a significant impact on drug discovery and new intervention development (Aban & George, 2015). In preclinical setup, interdisciplinary experiments produce an immense amount of information related to imaging, histology, pathology, gene sequence, morphology, etc.

The current work presented the implementation of DBMS in a preclinical research setting where data from different modalities were integrated. An interactive analytical tool was developed to analyze diffusion tensor imaging (DTI) metrics computed in a research study. The brief description of DTI is given below.

1.2 Diffusion Tensor Imaging

The diffusion tensor was originally proposed for use in magnetic resonance imaging (MRI) by Peter Basser in 1994 (Basser et al., 1994). DTI is one of the most popular noninvasive methods to characterize the dispersion pattern of water molecules in tissue. (Emsell et al., 2016). The diffusion of water molecules in tissues is restricted compared to free spaces.

However, the fiber architecture of tissues have a higher amount of diffusion along the fiber direction than in the perpendicular direction. So, DTI measurements can detect the structural integrity of neuronal fibers in the brain. Due to this reason, DTI has been used in studying various neurodegenerative pathologies including schizophrenia, traumatic brain injury, HIV, multiple sclerosis, autism, and aging (O'Donnell & Westin, 2011).

1.2.1 Principles of Diffusion Tensor Imaging

Diffusion is a process of thermally-driven displacement of water molecule due to collision with surrounding compartments. In the cerebrospinal fluid (CSF) the water movement is equal in all directions, which is isotropic, whereas the diffusion in white matter (WM) in the brain is not equal due to the presence of axonal membranes and myelin sheaths (Figure 1.6). This phenomenon is known as anisotropic nature (Alexander et al., 2007).

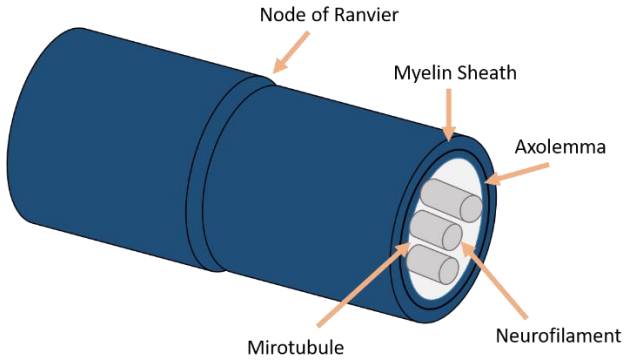


Figure 1.6: Axonal structure of white matter of the brain. The axon is covered by the myelin sheath with Ranvier nodes, which increases the conduction speed. Myelin sheath, axolemma, microtubules, and neurofilaments represent longitudinal axonal elements that could restrict the diffusion perpendicular to the axons than along with the axons.

1.2.1.1 Diffusion Weighted Imaging

Diffusion Weighted Imaging (DWI) is a standard MRI method to measure the diffusion profile of the water molecule at the three-dimensional voxel level. To generate DWI, the MRI sequences are made sensitive to the diffusion by adding two diffusion gradients to

the MRI sequence, normally T2 weighted echo planar imaging (EPI) sequence along the same directional axis (O'Donnell & Westin, 2011). The reason for choosing EPI spin echo is the sensitivity of the sequence to the water molecule. The two diffusion gradients are equal in magnitude and symmetrically centered around a 180º radiofrequency (RF) pulse as shown in figure 1.7. The amount of the diffusion in the direction of the applied gradient is quantified by a comparison of scans with and without diffusion sensitizing gradient. The first gradient is responsible for the phase shift, and the second one cancels the gained phase shift of non-moving stationary spins by reshasing (Huisman, 2010). Echo Time (TE) is the time from the center of the 90° pulse to the center of the echo.

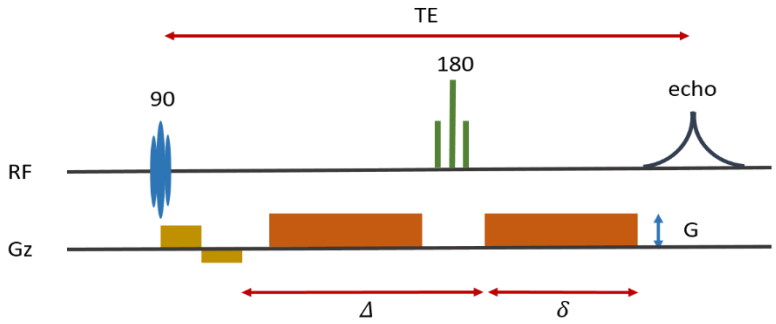


Figure 1.7: To produce a DWI, two equal magnitude gradients that are placed symmetrically centered around a 180 º RF pulse, are added to a regular MRI sequence.

The degree of diffusion weighting is also known as a dedicated b-factor (Equation 1)

$$b = \gamma^2 G^2 \delta^2 \left(\Delta - {}^{\delta}/_{3}\right) \quad (1)$$

Here, γ is the gyromagnetic ratio, G is the gradient amplitude, Δ is the pulse separation, δ is the pulse duration. Usually, two b-values are used to produce a meaningful interpretation. A higher degree of freedom for motion results in higher signal loss and appears dark in DWI, on the other hand, a low degree of freedom for motion appears

bright due to less MRI signal loss. The MR signal can be changed by changing *b*-value of the diffusion encoding gradient. And the b-value is changed by altering the strength G, the duration δ, or the time interval Δ. The resulting signal intensity is mapped as a two-dimensional image for each voxel in the DWI map (Lazar, 2010). A map of the diffusion coefficient D can be calculated by taking the difference between those two images where b-value is high and low. This map is called the apparent diffusion coefficient (ADC) map. The diffusion coefficient D is calculated from MR signal (Equation 2).

$$S = PD(1 - e^{-\frac{TR}{T_1}})e^{-\frac{TE}{T_2}}e^{-bD} \quad (2)$$

Here, S is the MR signal intensity in spin echo image, PD is proton density, T_1 is the longitudinal relaxation time, T_2 is transverse relaxation time, TR is repetition time, TE is the echo time, *b* is the diffusion-weighted factor (*b*- value) and D is the diffusion coefficient (representing the Brownian motion of water molecules) (Mori & Zhang, 2006).

The major advantage of ADC maps is that there is no effect of T_2 which may appear on DWI images. In the ADC map, the intensity of each pixel is proportional to the extent of diffusion. That means the brighter regions diffuse faster than darker regions, which is opposite to DWI map (Figure 1.8) (Mori & Zhang, 2006).

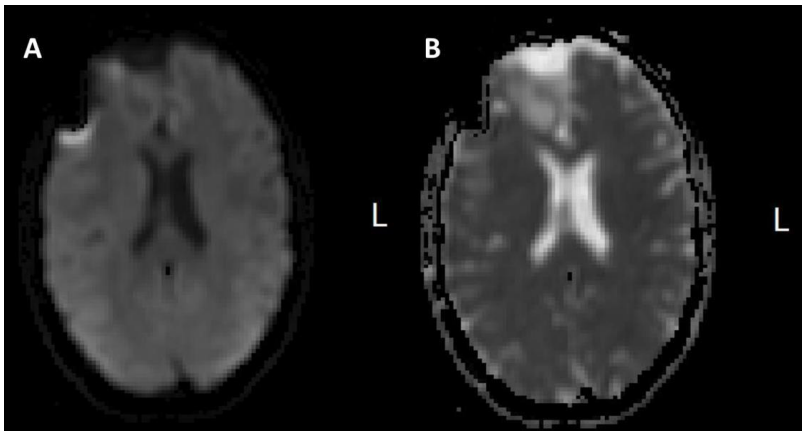


Figure 1.8: The difference between the DWI (A) and ADC (B) map where DWI showed faster diffusion for dark voxel and ADC showed faster diffusion for bright voxel (Mori & Zhang, 2006).

1.2.1.2 Diffusion Tensor Imaging

Diffusion Tensor Imaging is a diffusion MRI technique that uses positive definite 2nd order tensor to define the diffusion properties of water movement (Bihan et al., 2001). The diffusion tensor characterizes the situation where the Gaussian diffusions per unit of time differ in all directions. A 3 x 3 matrix described the displacement, where diagonal elements correspond to diffusivities along three orthogonal axes, and off-diagonal elements represent correlations between movements along these orthogonal axes (equation 3).

$$D\ (diffusion\ profile) = \begin{bmatrix} D_{xx} & D_{xy} & D_{xz} \\ D_{xy} & D_{yy} & D_{yz} \\ D_{xz} & D_{yz} & D_{zz} \end{bmatrix} \quad (3)$$

To construct the diffusion tensor, at least six non-collinear diffusion-weighted scans and one scan without diffusion weighting are required. As an ellipsoid diffusion tensor, the eigenvectors and eigenvalues describe the principal axes and lengths respectively, which are given by the diffusion distance in given time t. According to the Einstein equation (equation 4), the average squared displacements of molecules in the

sample (r^2) over time is proportional to the observation time t. Here, n = 6 refers to three dimensions.

$$r^2 = \mathrm{n}Dt \tag{4}$$

Squared displacement takes Gaussian form with peak positioned at zero movements and with the same equal probability of moving a given distance from the origin. The Einstein equation also suggests that the ellipsoid axes are scaled according to the square root of eigenvalues (Figure 1.9, 1.10). The orientation of the principal axes is characterized by three mutually orthogonal eigenvectors ($\varepsilon_1, \varepsilon_2, \varepsilon_3$) and with three eigenvalues ($\lambda_1, \lambda_2, \lambda_3$). The eigenvalues are called principal diffusivities and eigenvectors are principal directions of diffusion. The dominant fiber orientation in the voxel is parallel to the principal eigenvector (ε_1) associated with the largest eigenvalue (λ_1). The scalar invariants from diffusion tensor are calculated by using eigenvalues.

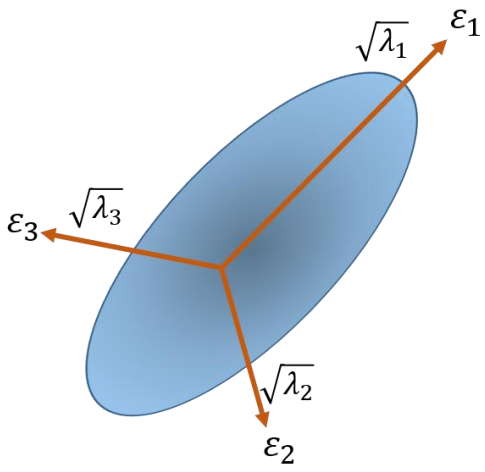


Figure 1.9: Diffusion tensor ellipsoid with the eigenvalues arranged by magnitude ($\lambda_1 > \lambda_2 > \lambda_3$).

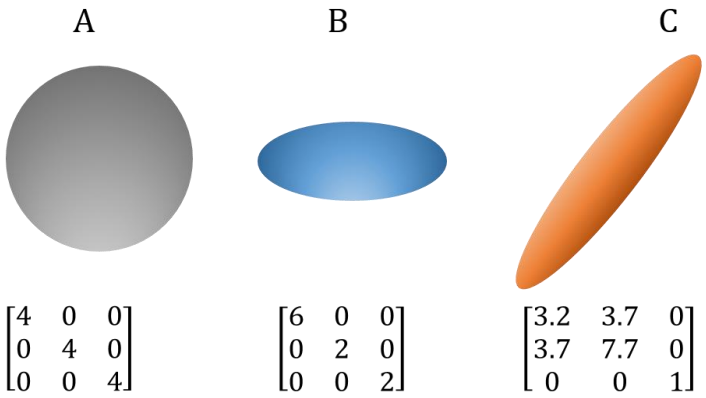


Figure 1.10: Tensor shapes and their diffusion 3 X 3 matrix. A) Isotropic diffusion, B) anisotropic diffusion, and C) rotated anisotropic tensor shape.

1.2.2 Scalar Invariants from the Diffusion Tensor

Some important scalar values can give an idea of the diffusion. Some of them are: Trace, Mean Diffusivity (MD), Fractional Anisotropy (FA), Relative Anisotropy (RA), etc.

1.2.2.1 Trace and Mean Diffusivity

Trace is the sum of the eigenvalues (equation 5). It is rotationally independent value. A related scalar value of trace is mean diffusivity (MD). MD is the average value of three eigenvalues (equation 6).

$$Trace = \lambda_1 + \lambda_2 + \lambda_3 \quad (5)$$

$$MD = \frac{Trace}{3} \quad (6)$$

MD can be decomposed into two components. One is axial (parallel, longitudinal diffusivity), which is $\lambda_{||} = \lambda_1$ and another one is radial diffusivity ($\lambda_{\perp} = \frac{(\lambda_2 + \lambda_3)}{2}$). MD is sensitive to the cellular abnormalities. The damaged tissue with increased diffusion shows higher MD compared with Radial Diffusivity (RD) and Axial Diffusivity (AD). MD is also known as ADC.

1.2.2.3 Fractional Anisotropy

Fractional Anisotropy (FA) is an important measure of DTI. It ranges from 0 to 1. For the non-directional tissues like CSF, this value is 0, whereas for WM fiber bundles such as in corpus collosum FA can be 0.9. FA (equation 7) describes the variation between the levels of diffusion measured in the different directions (Pierpaoli et al., 1996).

$$FA = \sqrt{\frac{3}{2}}\sqrt{\frac{(\lambda_1 - MD)^2 + (\lambda_2 - MD)^2 + (\lambda_3 - MD)^2}{{\lambda_1}^2 + {\lambda_2}^2 + {\lambda_3}^2}} \quad (7)$$

Where $\lambda_1$represents the principal eigenvalue parallel to the axonal axis, and λ_2, λ_3 characterize eigenvalues perpendicular to principal axon axis. FA defines the fiber integrity WM tissue in the brain by providing the level of directionalities. FA value reflects the change in WM due to the degree of myelination, axonal packing and axon size and coherence and co-linearity of fiber organization (Mori & Zhang, 2006). Multiple studies recommended that FA is associated with neurological and psychiatric conditions. It is also reported higher FA values in conditions related to WM disruption like in William syndrome, bipolar disorder, and Attention deficit hyperactivity disorder (ADHD).

1.2.2.4 Relative Anisotropy

Relative Anisotropy (RA) provides the comparison between the magnitude of the anisotropic part of diffusion tensor with the isotropic part by considering the ratio of the variance of the eigenvalues to their mean (equation 8) (Basser & Pierpaoli, 1996).

$$RA = \sqrt{\frac{(\lambda_1 - \lambda_2)^2 + (\lambda_1 - \lambda_3)^2 + (\lambda_2 - \lambda_3)^2)}{(\lambda_1 + \lambda_2 + \lambda_3)}} \quad (8)$$

1.2.3 DTI Data Processing

Bio-imaging core facility at the University of Nebraska Medical Center (UNMC) provides MRI imaging services including DTI data acquisition, processing, and analysis. Figure 1.11 shows the conventional process for MR- DTI data processing and analysis. After preparing the small animal (mouse or rat) for scanning, DTI data acquired by using 7T Bruker small animal scanner. The scanner generated image is needed to preprocess for analyzing the data. A custom computer program written in Interactive Data Language (IDL; Exelis Visual Information Solutions; McLean, VA, USA) for that purpose is used. The region of interest (ROI) extraction of computed DTI metrics results in a text-formatted file that includes the scalar quantities such as ADC and FA. A trained researcher reorganizes those values and merges with relevant animal information which includes animal name, animal group, scan date and respective ROI to create an Excel-formatted file to carry out the statistical analysis and produce graphical results. The newly created excel file is the only method of storing the data. Since the processing involves significant manual intervention, this may lead to possible human errors. Besides, it takes a few hours to process and compile the data for statistical analysis in an excel file.

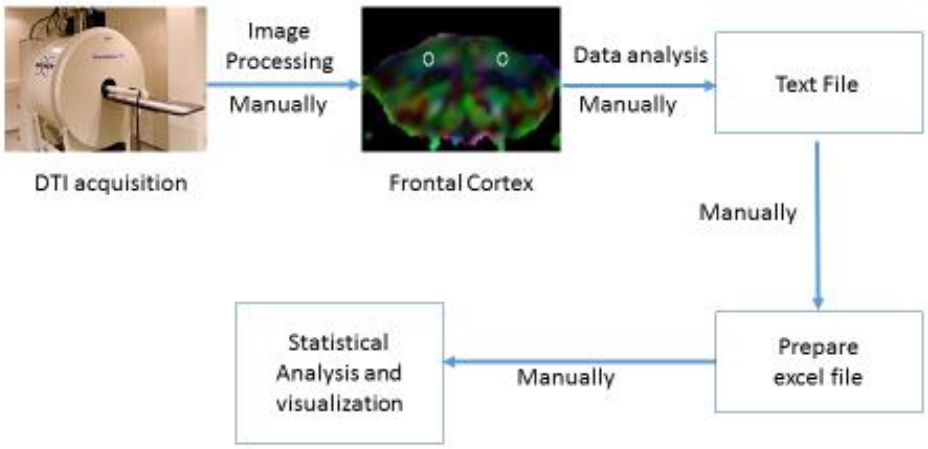


Figure 1.11: The current MR- DTI data processing and analysis workflow.

1.3 Structure of the Thesis

Taking into the account all tedious steps involved in the current data storage, processing and analysis pipeline, in this project, we propose to develop an entity-relational database and an interactive analytical tool (Figure 1.12). This computational tool will significantly reduce the processing time and human errors by automating several steps in the current data analysis pipeline. The analytical tool will convert multiple ROI text files into a CSV-formatted file that is compatible with the database format. The database will retain all the DTI data from different animals, which can be accessed from any computer in the university network. Statistical analysis (t-test for two group mean comparison, ANOVA for comparing mean within the group) and visualization of the dataset as plots (box plot, bar plot, and line plot) will be generated by the interactive DTI tool. For developing the relational database, MS-SQL server management studio 2014 (Express Edition) will be used. The interactive DTI analytical tool is developed by using R programming (an open source language), and R-studio (an open-source integrated development environment for R).

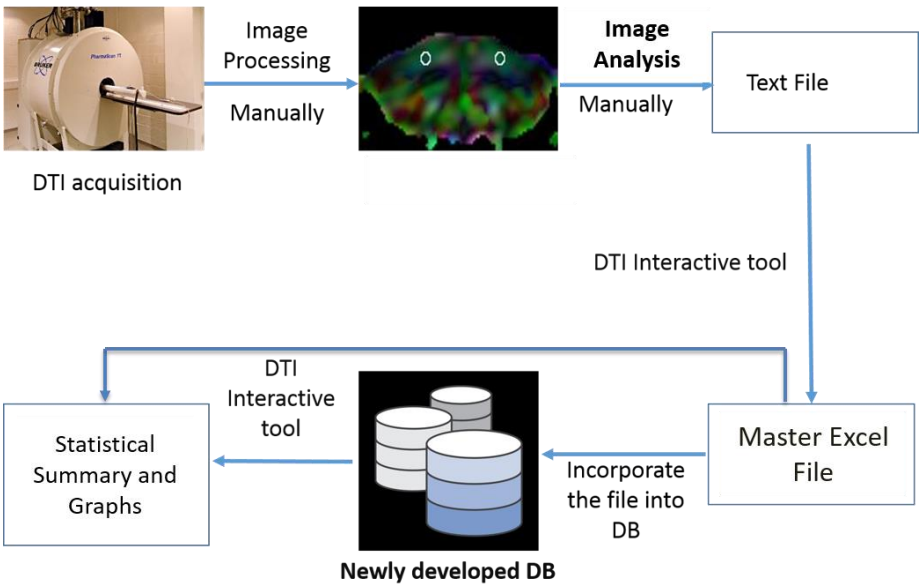


Figure 1.12: The proposed MR- DTI data processing and analysis workflow. The interactive DTI analytical tool converts the text- formatted file into a database compatible CSV-formatted file. After incorporating the file into the database, the tool can extract the data without going through the manual mapping and produce statistical summaries, graphs and plots interactively.

The present thesis work is focused on the development process of a biomedical database for preclinical imaging research and an interactive analytical DTI tool. In chapter 2, the methods and materials for developing the database and DTI tool are discussed. Chapter 3 presents the results from the interactive DTI analytical tool, and in the last chapter, the advantages of the new tool over the conventional process to analyze DTI data and future directions are discussed.

CHAPTER 2: METHODS AND MATERIALS

In the field of preclinical research, it is common to use multiple methods to acquire a wide spectrum of information to understand the basic biology of underlying diseases and to design new drugs and monitor the therapeutic efficacy. However, often it is quite challenging to combine and interpret the multimodal multidimensional data. This demands the development of relevant relational databases and tools for analyses. In this chapter, first, we discussed a database that could house data from different modalities and should be accessible as required. Then the details of designing and implementation of an interactive tool for analyzing DTI data were followed. The database has been developed by collaborating with Dr. Guda's Bioinformatics Laboratory at UNMC, and Navodita Upadhyay has helped to build the entity-relational database in Microsoft SQL Server.

2.1. Development of a Relational Database for DTI data

In this section, different components of the DTI relational database were discussed. The ERD, various data tables and the logical connections among them were presented. The following steps were carried out in building the present database.

2.1.1. Requirements Collection

The first step in the development of any database is to collect the requirements. For this purpose, all the processes involved in a preclinical biomedical imaging study from the MRI point of view were examined. In studying human diseases in animal models, sometimes animals may be injected with specific cells or virus after birth or at specific weeks of age. Studies may use animals of different models or strain. To check the disease progression, a trained technician collects blood samples from those animals to measure viral loads. The animal may be kept in different animal facilities of UNMC. Prior to scanning, a technician collects animals from the animal facility and returns to the same facility after

finishing the scanning. In some studies, before or during the scanning, the animal may be injected with nanoparticles to observe biodistribution of particles. Different types of scan protocols will be used to acquire data such as MRI and MRS. A trained technician processes and analyzes those data for formulating the results. Sometimes, it is required to sacrifice the animal after acquiring the data to perform histological experiments. Any animal that comes to the bio-imaging core belongs to a principal investigator (PI), and it also contains the Institutional Animal Care and Use Committee (IACUC) approval number. The IACUC number is issued to a PI by the research authority. Under one IACUC number, multiple studies can be conducted.

After considering all possible scenarios mentioned above, the data acquired through the pre, post and during scanning conditions, the conceptual model of the database was prepared. The database had 18 entities or tables in total (Table 2.1).

TABLE_CATALOG	TABLE_SCHEMA	TABLE_NAME	TABLE_TYPE
HMTS	dbo	ANIMAL	BASE TABLE
HMTS	dbo	BLEED	BASE TABLE
HMTS	dbo	COST_CENTER	BASE TABLE
HMTS	dbo	DESTINATION	BASE TABLE
HMTS	dbo	DISEASE	BASE TABLE
HMTS	dbo	DTI	BASE TABLE
HMTS	dbo	HISTOLOGY	BASE TABLE
HMTS	dbo	IACUC	BASE TABLE
HMTS	dbo	LC_MODEL	BASE TABLE
HMTS	dbo	MODEL	BASE TABLE
HMTS	dbo	MORPHOMETRY	BASE TABLE

HMTS	dbo	NANOMATERIALS	BASE TABLE
HMTS	dbo	PRINCIPAL_INVESTIGATOR	BASE TABLE
HMTS	dbo	QUEST_MODEL	BASE TABLE
HMTS	dbo	SCANTYPE	BASE TABLE
HMTS	dbo	SPECTRA	BASE TABLE
HMTS	dbo	STUDY	BASE TABLE
HMTS	dbo	TECH	BASE TABLE

Table 2.1: Entities of the preclinical database

2.1.2. Design of Tables in the Database

The **ANIMAL** table contains information about animals such as animal type, strain, gender, date of birth, scan date, housing cage information, disease type, etc. The **BLEED** table contains the records for each bleed at different time points. This table saves the information related to the date of bleed, and measurements from the blood panel tests, etc. The **COST_CENTER** table has the information about the cost center number for billing purposes. **DESTINATION** is the table for storing the information on the arrival and departure destinations of the animal before and after scanning. The **DISEASE** table stores disease name, disease date (if the disease is induced), comments, the source of disease, etc. **DTI** table stores the processed data of animals from different regions of the brain at multiple time points. **HISTOLOGY** table saves the results on brain region, stain type and density of animal tissues after the staining procedure. **IACUC** table contains the IACUC number issued to the individual PI by the animal research authority at UNMC. The **SPECTRA**, **LC_MODEL, QUEST_MODEL** tables store the results from spectral analysis including treatment, time points, animal IDs, area name, etc. The **MODEL** table contains information on the animal models. **NANOMATERIALS** table stores information on

different nanoparticles- name and suggested dosage as approved by NIH for different study purposes. **MORPHOMETRY** table consists of processed data from morphological experiments that include relaxation times analyses, signal intensity analysis, brain region, analysis type, brain volumetric changes, etc. The **TECH** table stores information on the technician who handles the animal. **PRINCIPAL_INVESTIGATOR** table stores information about the PI including the contact details. **SCANTYPE** contains information on different types of scans performed in the bio-imaging core facility. Finally, the **STUDY** table is for storing information related to various studies, which were designed and executed by our lab.

2.1.2. Logical Model Design

The next step in the relational database development was to design the logical model from the conceptual model. In the logical model, the relationship between entities was presented. For this purpose, the **ANIMAL** entity was considered as the center of our database because the activities in bio-imaging at all stages were related to the animals provided by the PIs. Also, all other non-imaging information was connected through the animal table.

Keeping this in consideration, all entities and attributes were created. As shown in the ERD in Figure 2.1, one table connected to another table with the primary key-foreign key relationship. To avoid cluttering, Figure 2.1 included only a few attributes, but the real entity contained a long list of attributes each represented a different experimental observation. To resolve the many-to-many relationships between IACUC and Principal_Investigator, a composite entity **"RESEARCH"** was created. The constraint names, parent entity and foreign entities for the relationship are shown in Table 2.2.

FOREIGN KEY TABLE	FOREIGN KEY COLUMN	FOREIGN KEY	PRIMARY KEY TABLE	PRIMARY KEY COLUMN	PRIMARY KEY NAME
Animal	Dest_Id	fk_animal_Dest_Id	Dest	Dest_ID	pk_Dest_Dest_ID
Animal	Dis_id	fk_animal_dis_id	Disease	Dis_ID	pk_Disease_Dis_ID
Animal	Stu_Name	fk_Animal_Stu_id	Study	Stu_Name	pk_Study_Stu_ID
Bleed	Animal_id	fk_bleed_animal_id	Animal	Animal_Id	pk_Animal_Animal_Id
Bleed	Tech_id	fk_bleed_tech_id	Tech	Tech_ID	pk_Tech_Tech_ID
Cost_Center	Stu_Name	fk_Cost_Center_Stu_id	Study	Stu_Name	pk_Study_Stu_ID
DTI	Animal_id	fk_DTI_animal_id	Animal	Animal_Id	pk_Animal_Animal_Id
DTI	Tech_Id	fk_DTI_Tech_id	Tech	Tech_ID	pk_Tech_Tech_ID
Histology	Animal_id	fk_Histology_animal_id	Animal	Animal_Id	pk_Animal_Animal_Id
IACUC	Stu_Name	fk_Iacuc_Stu_Name	Study	Stu_Name	pk_Study_Stu_ID
LC_Model	Animal_id	fk_LC_Model_animal_id	Animal	Animal_Id	pk_Animal_Animal_Id
Model	Animal_id	fk_Model_animal_id	Animal	Animal_Id	pk_Animal_Animal_Id
Model	Dis_id	fk_Model_dis_id	Disease	Dis_ID	pk_Disease_Dis_ID
Morphometry	Animal_id	fk_morphometry_animal_id	Animal	Animal_Id	pk_Animal_Animal_Id
Nanomaterials	Animal_id	fk_Nanomaterials_animal_id	Animal	Animal_Id	pk_Animal_Animal_Id

Nanomaterials	Tech_id	fk_Nanomaterials_tech_id	Tech	Tech_ID	pk_Tech_Tech_ID
Quest_Model	Animal_id	fk_Quest_Model_animal_id	Animal	Animal_Id	pk_Animal_Animal_Id
Research	Iacuc_Number	fk_Research_Iacuc_Number	IACUC	Iacuc_Number	pk_IACUC_Iacuc_Number
Research	PI_Id	fk_Research_PI_Id	Principal_Investigator	PI_ID	pk_Principal_Investigator_PI_ID
ScanType	Animal_Id	fk_ScanType_Animal_id	Animal	Animal_Id	pk_Animal_Animal_Id
Spectra	Stu_Name	fk_Spectra_Stu_Name	Study	Stu_Name	pk_Study_Stu_ID
Study	PI_ID	fk_Study_PI_ID	Principal_Investigator	PI_ID	pk_Principal_Investigator_PI_ID
tblDTI	Animal_id	fk_tblDTI_animal_id	Animal	Animal_Id	pk_Animal_Animal_Id
tblDTI	Tech_Id	fk_tblDTI_Tech_id	Tech	Tech_ID	pk_Tech_Tech_ID
tblLC_Model	Animal_id	fk_tblLC_Model_animal_id	Animal	Animal_Id	pk_Animal_Animal_Id
tblQuest_Model	Animal_id	fk_tblQuest_Model_animal_id	Animal	Animal_Id	pk_Animal_Animal_Id
tblSpectra	Stu_Name	fk_tblSpectra_Stu_Name	Study	Stu_Name	pk_Study_Stu_ID

Table 2.2: Relationship constraints of the preclinical database

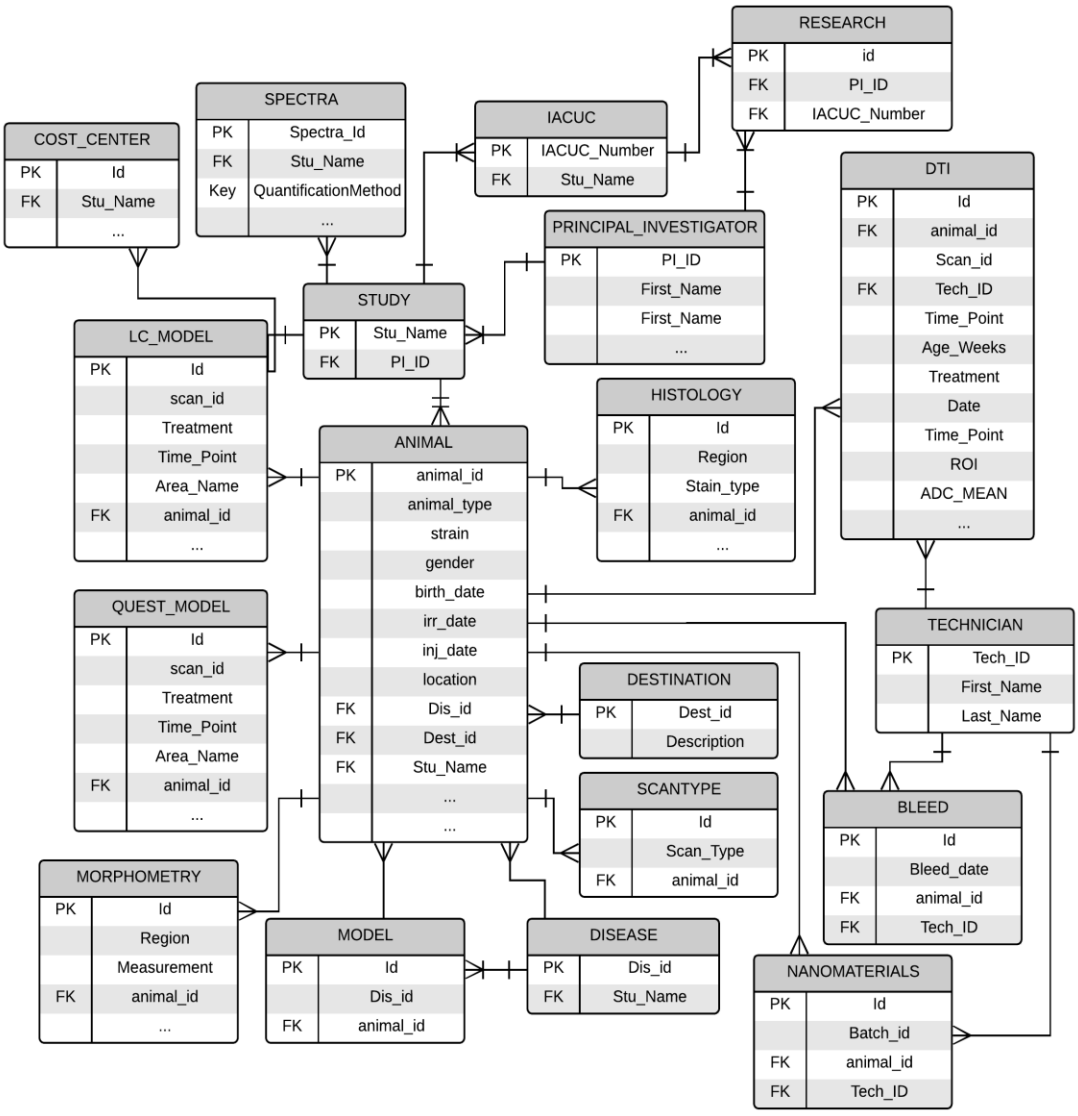


Figure 2.1: Entity Relationship diagram of the database.

2.1.3. Physical Model Design

The database was constructed using Microsoft SQL Server 2016 Express Edition. SQL Server helps to construct SQL Server relational databases in a systematic and user-friendly environment. SQL is a comprehensive database language that has statements for data definition, query and update. In addition, it has facilities for defining views on the database, for specifying security and authorization, for defining integrity constraints and for specifying transaction controls ("SQL", 2018).

The first step was to create data tables. The attributes were defined and the primary keys of each table were designated. Each entity was identified by a primary key. As an example, for the **ANIMAL** table, the *Animal_ID* attribute is the primary key that is unique for each animal. Data types and data lengths of each attribute were also defined while creating each table. Nineteen data tables were created for the current database.

The *Create Table* command was used to specify a new table by naming and specifying its attributes and constraints. The definition of a table can be changed by using the *Alter Table* command. The *Insert* command was used to add a single row to a table and the *Delete* command to remove rows from a table. The *Update* command was used to modify attribute values of one or more selected rows. SQL has one basic statement for retrieving information from a database, the *Select* statement. A list of database tables and attributes are given in Appendix A.

2.2. Design and implementation of an interactive tool for analyzing DTI measures

An important part of data analysis is data visualization, and it serves many different purposes. Visualization is useful for understanding the general structure and patterns of the data, particularly when analyzing large and complex research data sets, studying interdependence/effect between different parameters. Static plots, which do not allow interacting with graphics, may not be able to quickly present all information of data effectively. On the other hand, the interactive display provides flexibility and control to users. For this reason, an interactive tool for DTI data analysis and visualization was developed using R (Version 3.5.1), an open source programming language and R-studio (Version 1.1.456) an IDE of R. The interactive features of this tool were supported by different R- packages, which are mentioned in Table 2.3.

R- Package	Version	Purpose
shiny	1.1.0	Builds interactive web applications and dashboard with R
shinydashboard	0.7.0	Provides a theme on top of 'Shiny', making it easy to create attractive dashboards.
rJava	0.9-10	Low-level interface to Java VM. Allows creation of objects, calling methods and accessing fields.
ggplot2	3.0.0	A system for 'declaratively' creating graphics
RODBC	1.3-15	Communicates directly to the ODBC interface

xlsxjars	0.6.1	Collects all the external jars required for the xlxs package
shinyjs	1.0.0	Performs common useful JavaScript
stringi	1.1.7	Allows for fast, correct, consistent, portable, as well as convenient character string/text processing in every locale and any native encoding.
dplyr	0.7.6	A fast, consistent tool for working with data frame like objects, both in memory and out of memory.
Hmisc	4.1-1	Contains many functions useful for data analysis, high-level graphics, utility operations, functions for computing sample size and power, imputing missing values, advanced table making, variable clustering, and character string manipulation.
reshape2	1.4.3	Flexibly restructure and aggregate data
egg	0.4.0	Helps to customize 'ggplot2' objects.
lme4	1.1-18-1	Fit linear and generalized linear mixed-effects models.
multcomp	1.4-8	Simultaneous tests and confidence intervals for general linear hypotheses in parametric models, including linear, generalized linear, linear mixed effects, and survival models.

nlme	3.1-137	Fit and compare Gaussian linear and nonlinear mixed-effects models.

Table 2.3: List of R- packages used to build the interactive tool

The interactive dashboard was designed in a uniform style with four tabs, each for a different functionality. These functionalities include: formatting and converting user provided dataset into comma-separated value (CSV) format document, connecting to the newly built database to extract the desired dataset for analysis, visualizing the data for qualitative assessment, and generating a summary of the statistical analysis. The four tabs are named 1) Data Formatting, 2) Data Extraction, 3) Data Visualization, and 4) Statistical Analysis (Figure 2.2). A brief description of the functionality of each tab is given below.

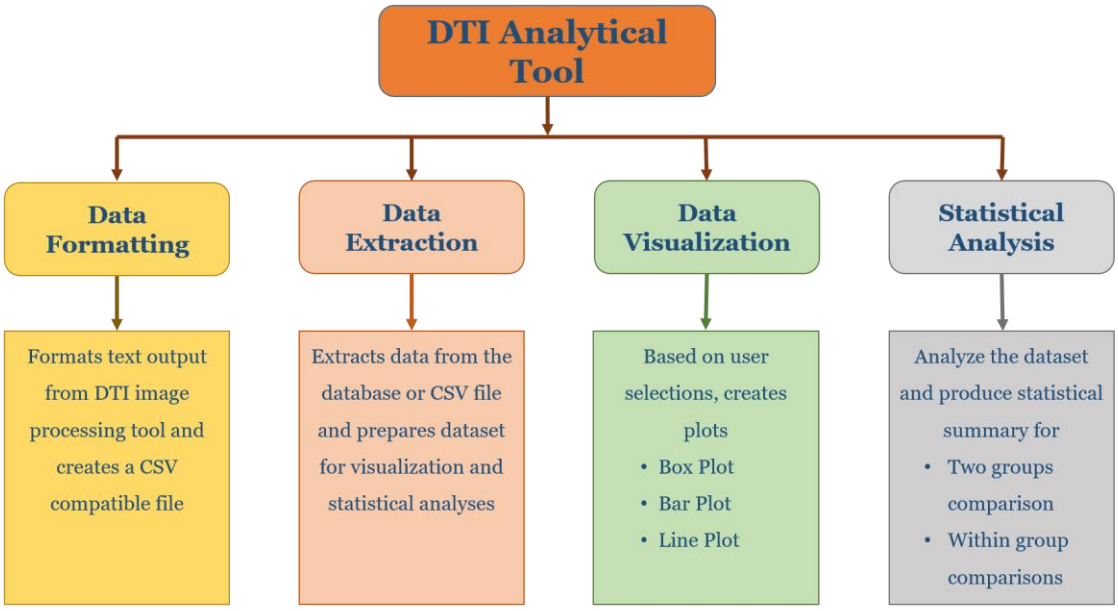


Figure 2.2: Schematic representation of the four tab sets used in the interactive DTI analysis tool.

2.2.1. Data Formatting Tab Set

“Data Formatting” is the first tab set on the interactive tool. The purpose of this tab set was to prepare a database-compatible dataset without manual intervention. The DTI data would, normally, be processed and the quantitative diffusion measures extracted from the region of interest (ROI) using the tools outside of the database. The conventional output from the DTI processing tools is a text-formatted file without any information related to the animal and procedure. In order to insert these datasets appropriately into the database, these files should have information related to the animal and scan such as Animal ID, Scan ID, Technician ID, Age of the animal in weeks, Group name of the animal (Control, Treated, Infected, etc.), Scan date, and data Time Point with relevant ROI name and quantitative DTI value.

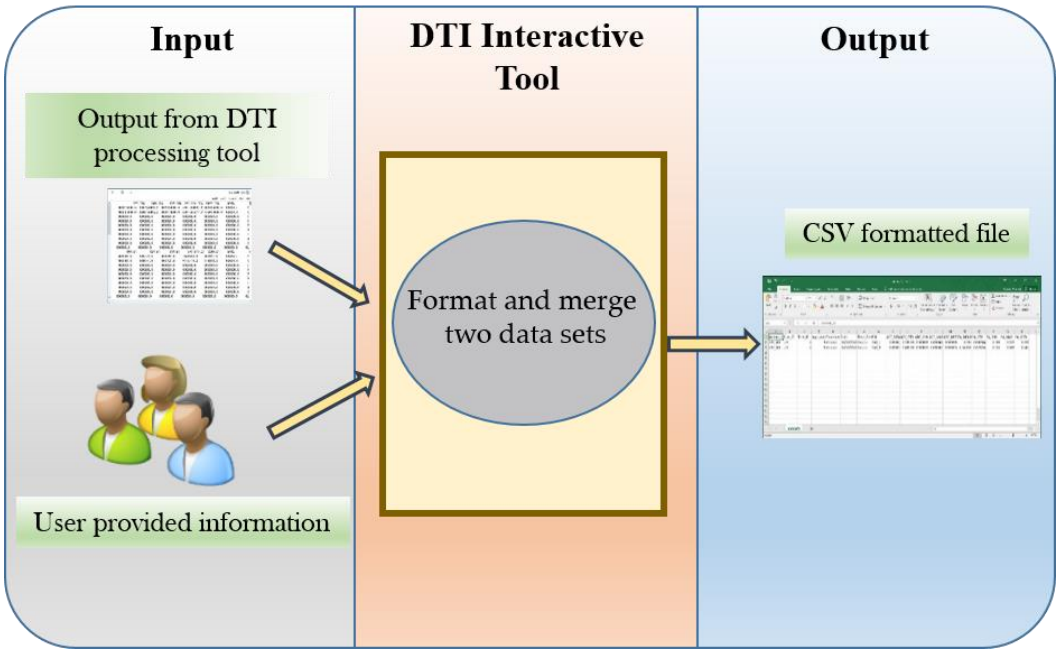


Figure 2.3: Steps in Data formatting process.

As shown in Figure 2.3, there were two steps involved in data formatting process. In the first step, the user needed to input the information regarding the animal in the animal

information form and load the text-formatted file from the local computer. The second step was to format and merge the information that was provided by the user and produce a CSV formatted database compatible file. The CSV output file had eighteen columns: “Animal_ID”, “Scan _ID”, “Tech_ID”, Age_Weeks”, Animal_Group”, “Date”, Time_Point”, “ROI”, “ADC_MEAN”, “ADC_STD”, “ADC_MIN”, “ADC_MAX”, “ADC_MED”, “FA_MEAN”, “FA_STD”, “FA_MIN”, “FA_MAX”, “FA_MED”. For each ROI, two rows of information were generated. One row was for the information on brain left side ROI and another was for the right side ROI. For example, if we inserted the ROI value “CA2”, the first row ROI value would be “CA2_L” and second row would be “CA2_R”. The Column names started with “ADC” and “FA” carried the value of anisotropic diffusion coefficient and fractional anisotropy respectively from DTI calculations. In the data insert form, “Animal ID”, “Animal Group”, “Time point” and “load the text file” were mandatory fields as without this information it was not possible to insert data in the appropriate location in the database. After incorporating the information into those fields, the “Update table” button would be enabled. Clicking this button would merge two sets of information and produce a CSV formatted file and would be displayed in the Data presentation area of the interactive tool (Figure 2.4).

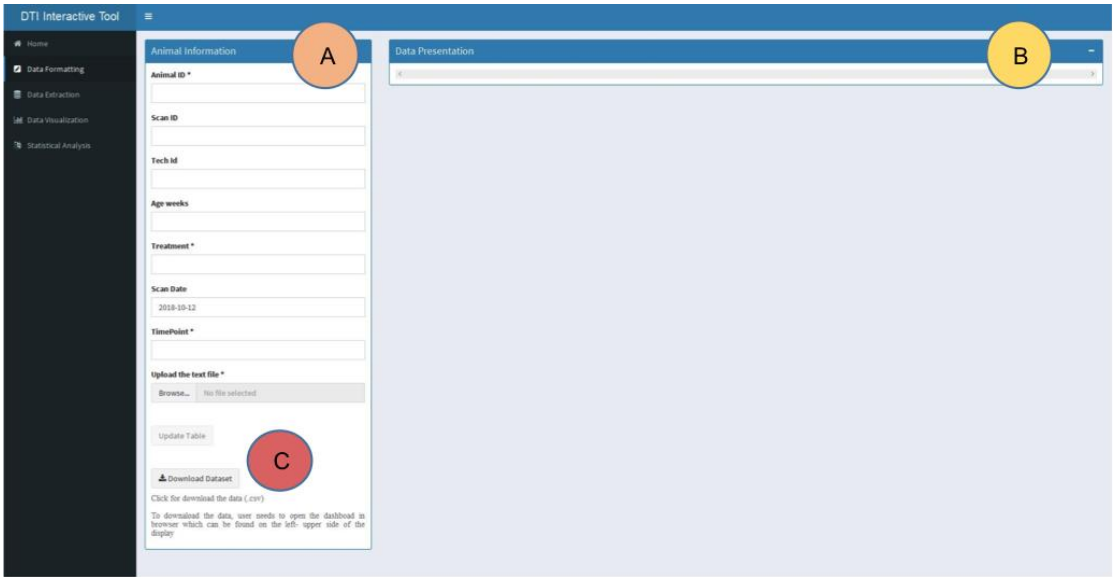


Figure 2.4: “Data Formatting” tab set of the DTI interactive tool. A) Form for incorporating animal information and text formatted file, B) the display box for merged data set, C) Download button for saving the CSV formatted dataset.

The R packages, *dplyr* and *reshape2* were used to format and merge two sets of information. “Download Dataset” button allows the user to download the dataset to the local disk for storing purpose. The saved file had the same format and data type as DTI table in the database.

2.2.2. Data Extraction Tab Set

“**Data Extraction”** had the functionality to communicate with the database and extract columns and rows from the DTI table. There were two types of extractions from the database. One was default extraction and another was a selective extraction. The default extraction provides information of “Animal_ID”, “Treatment”, “Time_Point”, “ROI”, “FA_MEAN”, “ADC_MEAN”, and “RA_MEAN” from the database, and the selective extraction was a subset of default extraction based on the user provided animal name. Each row of the extracted data list corresponds to one time point of an individual animal. It calculated the average value of “FA_MEAN”, “ADC_MEAN”, and “RA_MEAN” from left

and right ROI values. The selective extraction was needed to create a data set for visualization. The user would need to create a data set by selecting “Animal Name” involved in any specific study. Both the “Default Extraction” and “Selective Extraction” had the functionality to download extracted datasets (Figure 2.5).

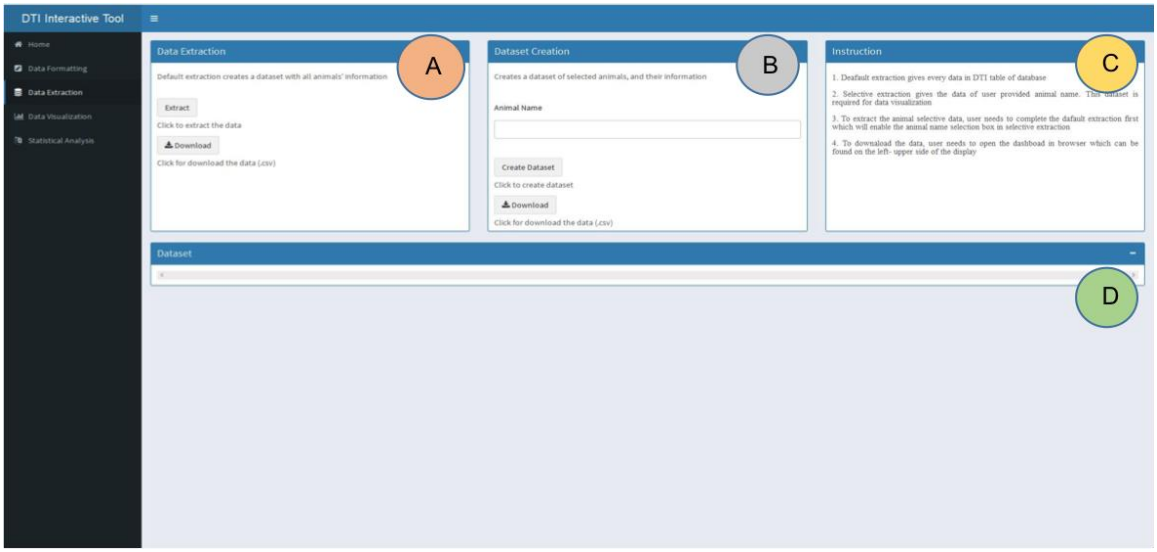


Figure 2.5: “Data Extraction” tab set of the DTI interactive tool. A) Default Data extraction, B) Selective Data extraction, C) Instructions for the user, and D) Box for displaying the extracted data.

The instructions on how to use each tab set was also provided in a separate box. To establish the communication between the interactive tool and database, an ODBC connection was created for MSSQL Server 2016 by providing the server information. And a script was written in R to connect to the database. *RODBC* package was used to build the communication between the tool and database.

2.2.3. Data Visualization Tab Set

“**Data Visualization**” produces box plot, bar plot, and line plot to visualize the DTI dataset based on different options, determined by the user. There were five subtab sets in “Data Visualization”: 1) Data Source Selection, 2) Plot Dataset Creation, 3) Box plot, 4) Bar plot,

and 5) Line plot. In “Data Source Selection” subtab, the user was able to choose the data source for creating a dataset for visualization. There were two options. One was for external CSV file, which might be downloaded from “Data Extraction” tab set. Another one was “DB Based”, which was the result of “Selective Data Extraction”. In that case, the user was not required to download the dataset. “Data Source Selection” also had a display box to see the loaded data from the source. A box with the instructions for the user was also included. Subtab “Plot Dataset Creation” had four mandatory fields to insert information. They were “Select Animal Group”, “Select Time Point”, “Select ROIs”, and “Select DTI metrics” (Figure 2.6). Those fields’ options would change dynamically with the dataset. The “Box plot”, “Bar plot”, and “Line Plot” subtab panel were for visualizing the output.

Figure 2.7, describes the steps involved in interactive DTI analytics tool. . After selecting the data by user, the dataset was prepared for visualization and statistical analysis using *reshape2, dplyr, and Hmisc* packages. The newly created dataset had nine columns: 1) “Treatment”, 2) “Time_Point”, 3) “ROI”, 4) “variable”, 5) “N”, 6) “value”, 7) “sd”, 8) “se”, and 9) “ci”. Here, “Treatment” defined the animal group of study, “Time_Point” was for the specific time point in the study, “ROI” was for the region of interest of the brain, DTI metrics’ name in the column “variable”, “N” was the number of animals at a specific time point, “value” was the calculated average value of DTI metrics of a specific ROI at a specific time point, “sd” was the standard deviation within the group, “se” was the standard error within group and “ci” was the confidence interval.

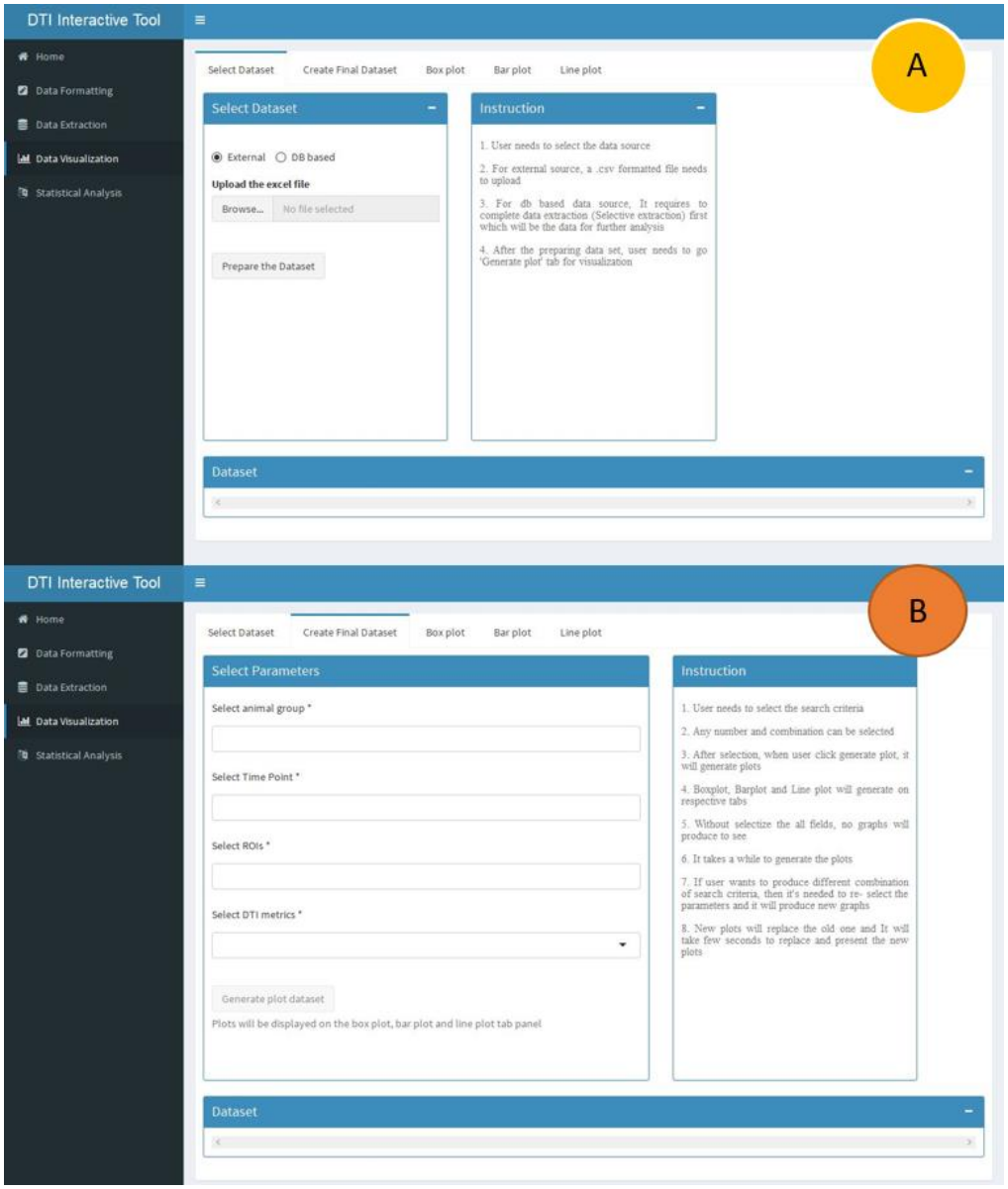


Figure 2.6: “Data Visualization” tab set of the DTI interactive tool. A) “Select Data Source” subtab set, B) “Plot Dataset Creation” subtab set.

After completing the data wrangling (Figure 2.7 A), that dataset was used to generate plots. R packages, *ggplot2 and egg,* were used to produce box, bar, and line plots. Box plot was used to check for the outliers in the data, and bar plot and line plot were to observe the changes in DTI metrics at different time points for different groups

and ROI's. Plots were displayed in respective subtab set. "Box plot", "Bar plot" and "Line plot" had "Download" button to download the plots in jpeg image format.

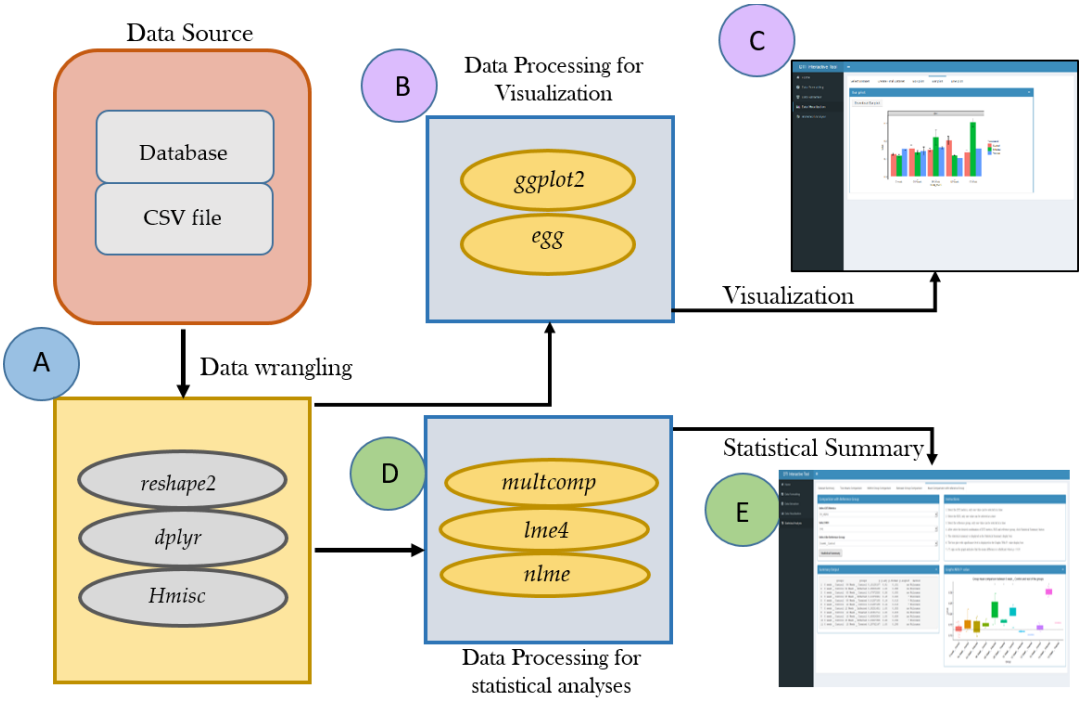


Figure 2.7: Architecture of Data visualization and statistical summary. A) Dataset preparation for visualization and statistical analysis, B) Data processing for visualization, C) Output of visualization, D) Data processing for statistical analyses, E) Output of statistical summary.

2.2.4. Statistical Analysis Tab Set

The last tab set of the interactive tool was **"Statistical Analysis"**. The **"Statistical Analysis"** tab set had "Dataset Summary", "Two-means comparison", "Within-group comparison", "Between Group Comparison", and "Mean Comparison with Reference Group" as five subtab sets (Figure 2.8). "Dataset Summary" described the dataset by providing the number of samples at each time point, number of ROIs, DTI metrics and their names. The subtab set "Two-means comparison" compared the mean of DTI values

of an ROI between different animal groups and time points. The flow chart for selecting the suitable statistical method for two mean comparisons is presented in Figure 2.9.

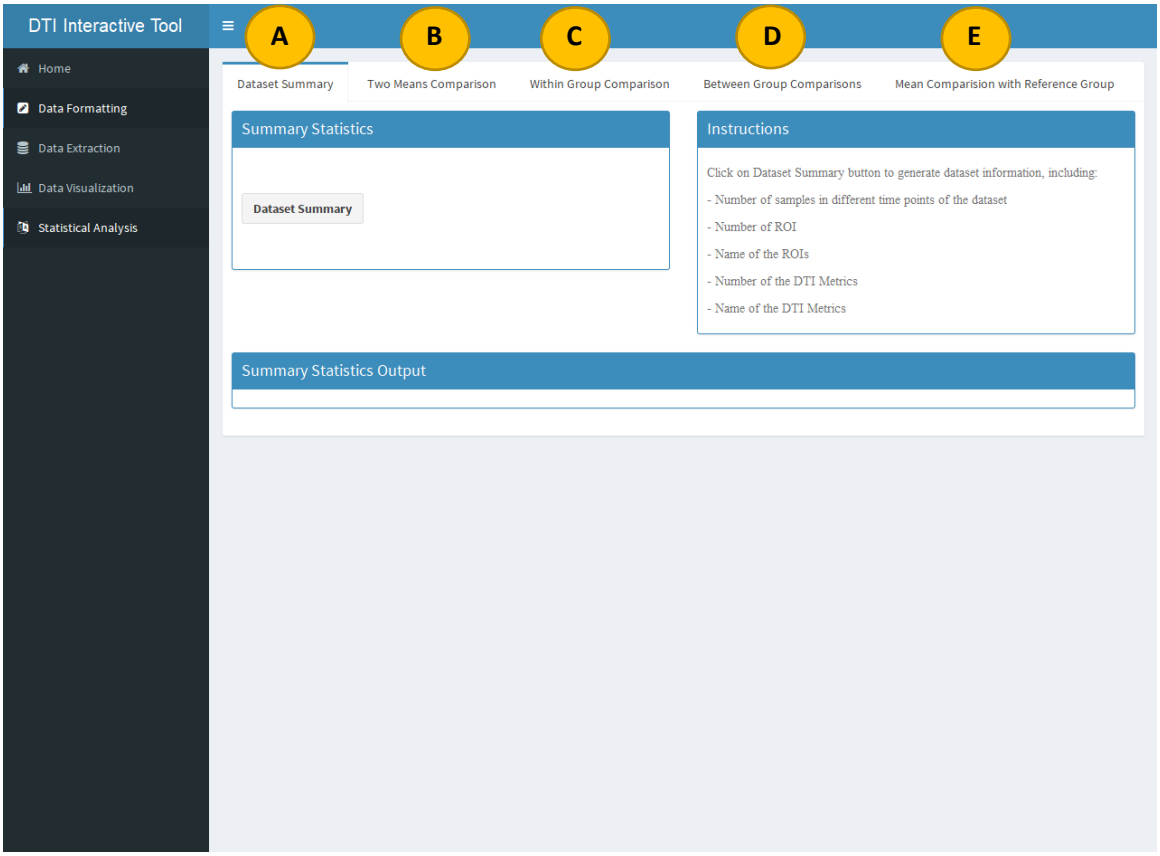


Figure 2.8: “Statistical Analysis” tab set. A) Dataset Summary subtab set, B) Two means comparison subtab set, C) Within group comparison subtab set, D) Between the group comparisons subtab set, E) Mean comparison with reference group subtab set.

The first step was to determine between parametric and non-parametric tests for statistical analysis. For this, the sample size at each time point was considered. If this number was greater than 20, a parametric test was performed, otherwise, a non-parametric test was performed *(Corder & Foreman, 2009).* For the parametric test, two-sample t-test and for the non-parametric test, Wilcoxon rank sum test with continuity correction was applied. Before performing the final statistical analysis, it was required to check if paired or non-paired comparison needed to perform between the two groups. If

the statistical analysis was between the same animal group at different time points, and both time points had the same number of observation then paired group analysis was performed else considered both groups as non-paired. “Within-group comparison” provided the comparison of the mean of the same group at all-time points. A repeated measure of ANOVA was selected for performing the within-group comparison. “Between-Group Comparisons” provided mean differences among the groups in each time points. Each group was considered as an independent variable in this analysis. A p-value of 0.05 was used to determine if the mean change was significant or not. In “Mean Comparison with Reference Group”, DTI metrics of all groups were compared with a user selected reference group. In “Between-Group Comparisons”, DTI metrics was compared among the groups in each time point. For between the group comparisons and mean comparison with the reference group, Wilcoxon signed rank test was performed to generate the statistical summary.

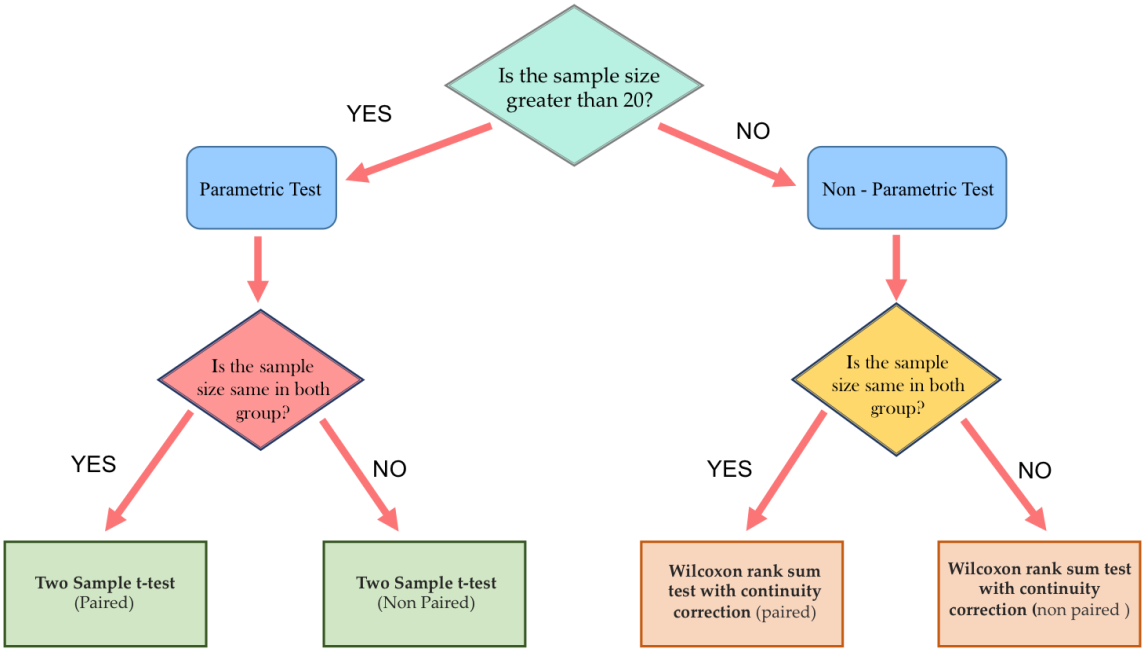


Figure 2.9: Decision tree for choosing the appropriate statistical analysis method.

As shown in Figure 2.7, the dataset prepared to visualize the graphs and plots and to perform the statistical analysis (Figure 2.7 A) was used. The R packages, *multicomp, lme4, and nlme* were used to generate the statistical summary report. Every subtab set had a summary output box to present the results.

2.2.5 Quality Assurance of the DTI Interactive Tool

Quality assurance is an integrated part of software development. To ensure the performance of the developed DTI interactive tool, the quality assurance testing was performed. The method of testing was manual and the functional testing methods were followed. For this purpose, test cases based on the functionality of the four tab sets of DTI interactive tool were written as shown in Table 2.4.

Sl No	Summary	Prerequisite	Test Case	Expected results
1	User can generate a csv formatted file by merging the animal information form and uploaded text file	1. Computer should be connected to the internet connection 2. R program should be running from R- Studio	1. User clicks on the "open in browser" on the upper left corner of the tool.	1. Tool opens in browser.
			2. User needs to fill the mandatory fields in the "Data Formatting" tab set which includes "Animal ID "Treatment", "Time_Point", "upload the text file"	2. The "Update Table" button is enabled
			3. User clicks on the "Update Table" button	3. The merged dataset is displayed on the "Data Presentation" box
			4. User clicks on the "Download" button	4. A download window opens up.
			5. User saves the file in the desired folder	5. The file is saved in the user provided location

2	User can extract and download data from the database by using the interactive tool	1. Computer should be connected to the internet connection 2. R program should be running from R- Studio 3. A connection between the database and the tool should be established	1. User clicks on the "open in browser" on the upper left corner of the tool.	1. Tool opens in browser.
			2. User clicks on "Data Extraction" tab set	2. Data extraction window is open which has two box, one for default extraction and another for selective extraction
			3. User clicks on the "Extract" button	3. A dataset with all animal information in DTI table from the database is extracted and displayed on "Dataset" display box and in the "Selective Extraction", the unique value of the animal names are available for selection in the "Animal Name" field.
			4. User clicks on the "Download" button	4. A download window opens up.
			5. User saves the file in the desired folder	5. The file is saved in the user provided location

			6. User selects the animal names from "Animal Name" field in "Selective Extraction" and clicks the "Extract Dataset"	6. A dataset is generated with the provided animal names' information and replaced the dataset from default extraction.
			7. User clicks on the "Download" button	7. A download window opens up.
			8. User saves the file in the desired folder	8. The file is saved in the user provided location
3	User can generate box plot, bar plot, line plot and able to download those plots	1. Computer should be connected to the internet connection 2. R program should be running from R- Studio 3. A connection between the database	1. User clicks on the "open in browser" on the upper left corner of the tool.	1. Tool opens in browser.
			2. User clicks on "Data Visualization" tab set	2. Data visualization window opens which has five subtab sets. They are "Dataset Selection", "Plot Data Creation", "Box Plot", "Bar Plot", and "Line Plot".

		and the tool should be established 4. Steps involved in data extraction should be completed	3. User needs to select the data source. If the data source is external then it's needed to load a CSV formatted file, which has the same column structure as the output of "Selective Extraction". If the source is "DB Based" then it's automatically used the result of "Selective Extraction".	3. "Prepare Dataset" button becomes enable
			4. User clicks on the "Prepare Dataset" button	4. Dataset appears on the "Dataset" display box and selection fields values on "Plot Data Creation" become available for selection
			5. User click on "Plot Data Creation"	5. "Plot Data Creation" subtab set opens
			6. User needs to fill the mandatory fields "Select Animal Names”,	6. The "Generate Plot Dataset" button becomes active for next step.

			“Select Animal Group", "Select ROI", and "Select DTI Metrics"	
			7. User clicks on the "Generate Plot Dataset" button	7. A dataset is generated based on the selection, and box plot, bar plot and line plot are generated
			8. User goes to the "Box Plot" subtab set	8. Box plot is displayed based on the input of the user on "Plot Data Creation" step.
			9. User goes to the "Bar Plot" subtab set	9. Bar plot is displayed based on the input of the user on "Plot Data Creation" step.
			10. User goes to the "Line Plot" subtab set	10. "Line plot is displayed based on the input of the user on "Plot Data Creation" step.
4	User can produce and download the statistical summary,	1. Computer should be connected to the internet connection	1. User clicks on the "open in browser" on the upper left corner of the tool.	1. Tool opens in browser.

	which includes dataset summary, two mean comparison, within group comparison of mean, mean comparison between the group, and mean comparison with a reference group	2. R program should be running from R- Studio 3. Steps involve in "Data set Creation" on "Data Visualization" tab set should be completed	2. User clicks on "Statistical Analysis" tab set	2. "Statistical Analysis" window opens up which has three subtab set. They are "Dataset Summary", "Two Mean Comparison", "Within Group Comparison", “Between Group Comparisons”, and “Mean Comparison with Reference Group”.
			3. User clicks on "Dataset Statistical Summary"	3. Dataset summary generates and presents in "Dataset Summary" display box and selection fields values on "Two mean Comparison" and "within Group Comparison" become available for selection
			4. User clicks on “Download Summary”	4. A text formatted summary generates.
			5. User clicks on the "Two Mean Comparison" subtab set.	5. "Two Mean Comparison" subtab set opens

			6. User selects the values for all selection fields from available options and clicks on "Generate Statistical Summary" button in "Two Mean Comparison" subtab set	6. The summary shows on "Summary Output" display box
			7. User clicks on “Download Summary”	7. A text formatted summary generates.
			8. User clicks on the "Within Group Comparison" subtab set.	8. "Within Group Comparison" subtab set opens
			9. User selects the values for all selection fields from available options and clicks on "Generate Statistical Summary" button in "Within Group Comparison" subtab set	9. The summary shows on "Summary Output" display box

			10. User clicks on “Download Summary”	10. A text formatted summary generates.
			11. User clicks on the "Between Group Comparison" subtab set.	11. "Between Group Comparison" subtab set opens
			12. User selects the values for all selection fields from available options and clicks on "Generate Statistical Summary" button in "Between Group Comparison" subtab set	12. The summary shows on "Summary Output" display box
			13. User clicks on “Download Summary”	13. A text formatted summary generates.
			14. User clicks on the "Mean Comparison with Reference Group" subtab set.	14. " Mean Comparison with Reference Group " subtab set opens

			15. User selects the values for all selection fields from available options and clicks on "Generate Statistical Summary" button in the "Mean Comparison with Reference Group" subtab set	15. The summary shows on "Summary Output" display box
			16. User clicks on “Download Summary”	16. A text formatted summary generates.

Table 2.4. Test cases used to test the accuracy and robustness of the interactive tool.

2.3. Test Dataset Creation

The functionalities of the interactive tool were tested by performing exploratory analysis by using a test dataset consisted of DTI data acquired on 15 mice in traumatic brain injury (TBI) study. Those mice were grouped into "**Group A**" (n = 5), "**Group B**" (n = 6), and "**Group C**" (n = 4).

2.3.1. Experimental Study Design

Data were acquired longitudinally at five time points (Pre-scan/0 week, after TBI at 4 weeks, 8 weeks, 12 weeks and 15 weeks) for "**Group A**", "**Group B**", and "**Group C**". "0 week" considered as pre-scan (or baseline scan) for all three groups. After acquiring the data at "0 week", TBI was created in "**Group B"** and "**Group C"**. "**Group B"** was started getting treatment drug after 4 week's scan. The data acquisition was continued for **"Group A"**, **"Group B"** and **"Group C"** at 8 weeks, 12 weeks and 15 weeks from the baseline scan. Due to the disease progression, a few mice died during the study. Three DTI metrics; ADC, RA, and FA were measured. These measurements were obtained from 12 regions of the brain (Frontal Cortex, Corpus Collosum, Corpus Collosum2, Cerebral Cortex, Cerebral Cortex (M2), Whisker Barrels, CA1 (Hippocampal subfield) , CA3 (Hippocampal subfield), Dentate Gyrus, Substantia Nigra, Medulla, Cerebellum, Hippocampus). The dataset consisted of 32 columns and 1924 rows. In figure 2.10, the timeline of this study is shown.

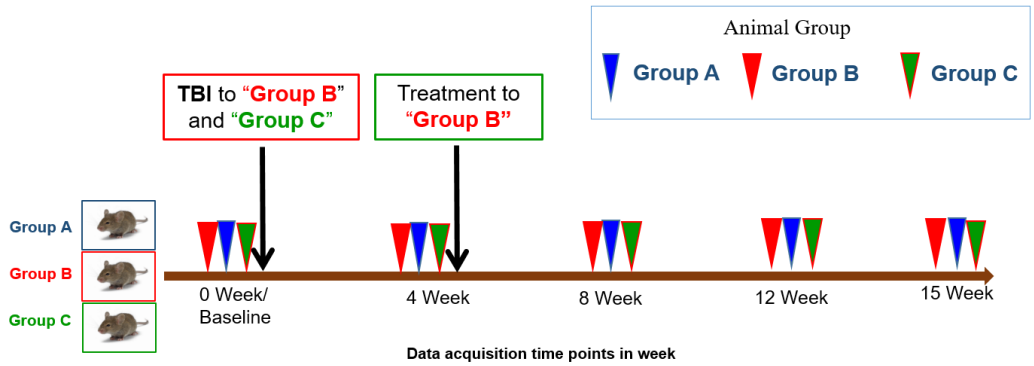


Figure 2.10: Timings for data acquisition. After 0 week's scan, TBI was created in "Group B" and "Group C" and after 4 week's scan, "Group B" started getting treatment drug.

Description of the DTI data analyses was discussed below.

2.3.2. DTI Data Processing

A DTI processing custom computer program written in IDL was used to preprocessed and quantify the diffusion-tensor data. Preprocessing involved checking and removing the corrupted dataset. In order to perform quantitative analysis, maps of the tensor diffusivities (λ_1, λ_2, and λ_3), ADC and FA (Figure 2.11 A) were generated. The output of the DTI metrics analyses was a text-formatted file (Figure 2.11 B). Then, those results were rearranged for preparing excel- formatted file to carry out the statistical analyses and produce the results in graphs.

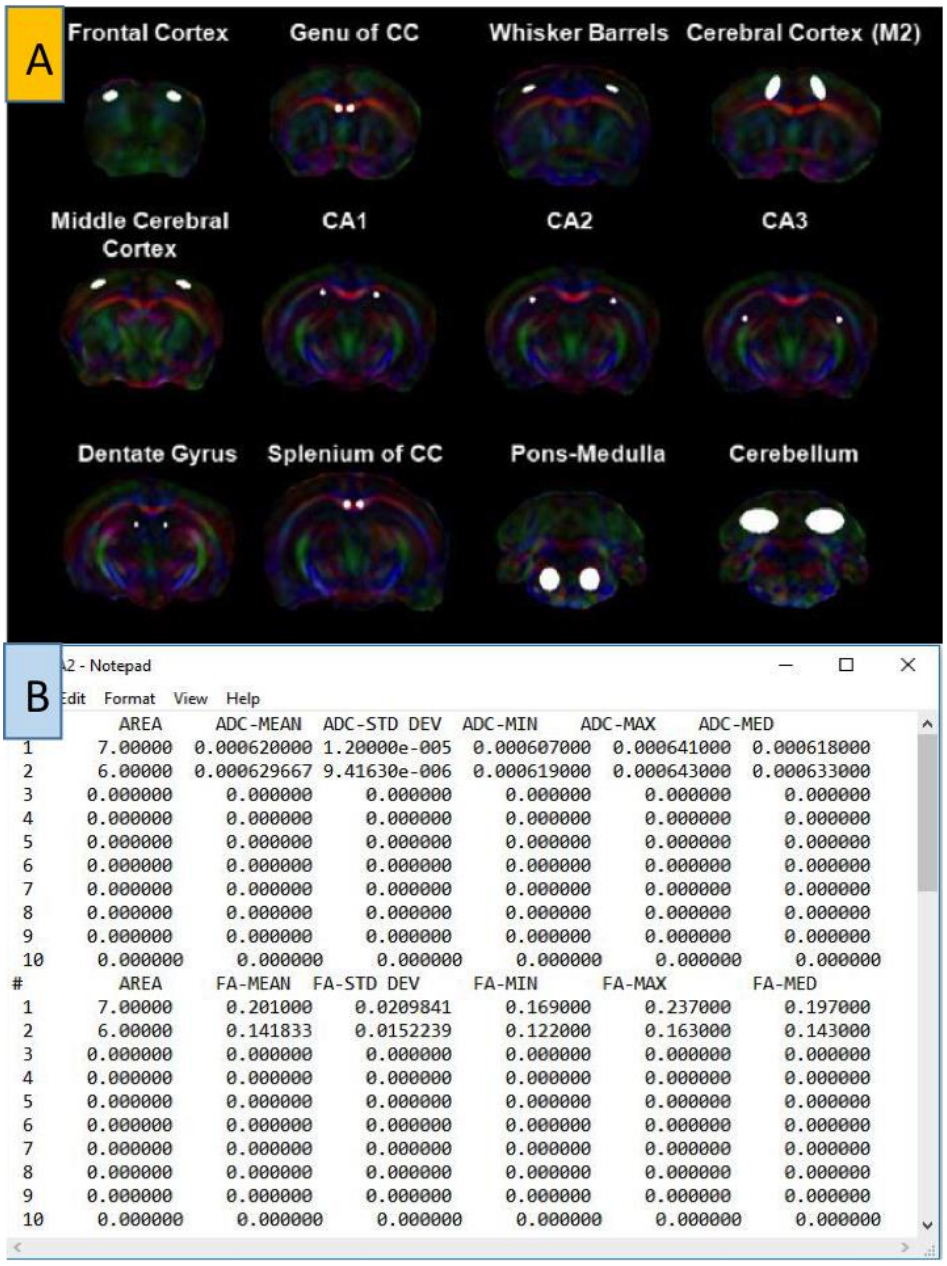


Figure 2.11: DTI data analyses output. A) DTI image after processing (Boska et al., 2014), B) Text-formatted data processing output for one ROI.

In this chapter, different steps involved in the development of the biological entity-relationship database and the DTI interactive analytical tool were described. The accuracy and the robustness of the tool is discussed in the next chapter.

CHAPTER 3: RESULTS

Software Quality Assurance or simply software testing is an essential part of Software Development Cycle that verifies the accuracy and robustness of the developed software in various test case scenarios. In this chapter, a step-by-step assessment of the functionalities of the developed interactive DTI tool is demonstrated using the DTI test dataset. The description of this test data is presented in section “2.3. Test Data Creation” of Chapter 2. The tool is tested following the test cases mentioned in subsection “2.2.5. Quality Assurance of the DTI Interactive Tool” of Chapter 2. Each case of the test-case scenario is evaluated and the results are documented. To verify the correctness of the present DTI analytical tool, the produced results are compared with results from the Statistical Analysis Software (SAS) using the same test dataset with similar use cases.

The first case of testing is to verify the functionality of the “Data Formatting” tab set. The expected result of this step is to convert the text-formatted DTI result of individual ROIs into a database compatible CSV file for data storing purpose. The CSV file should have the DTI values of ROIs’ and relevant information about the animal and data acquisition. Figure 3.1 shows the output of the merged information provided by the user to DTI text file. “HIV_HC_301” animal was used for creating the dataset. Information related to the animal and the text file of the ROI was also incorporated. To add another “ROI” data of the same animal, a new “ROI” text file was uploaded. The interactive tool read the title of that file and combined the new rows with existing rows of the same animal data. Every time a new ROI was added, the updated dataset was displayed on the “Data Presentation” display box.

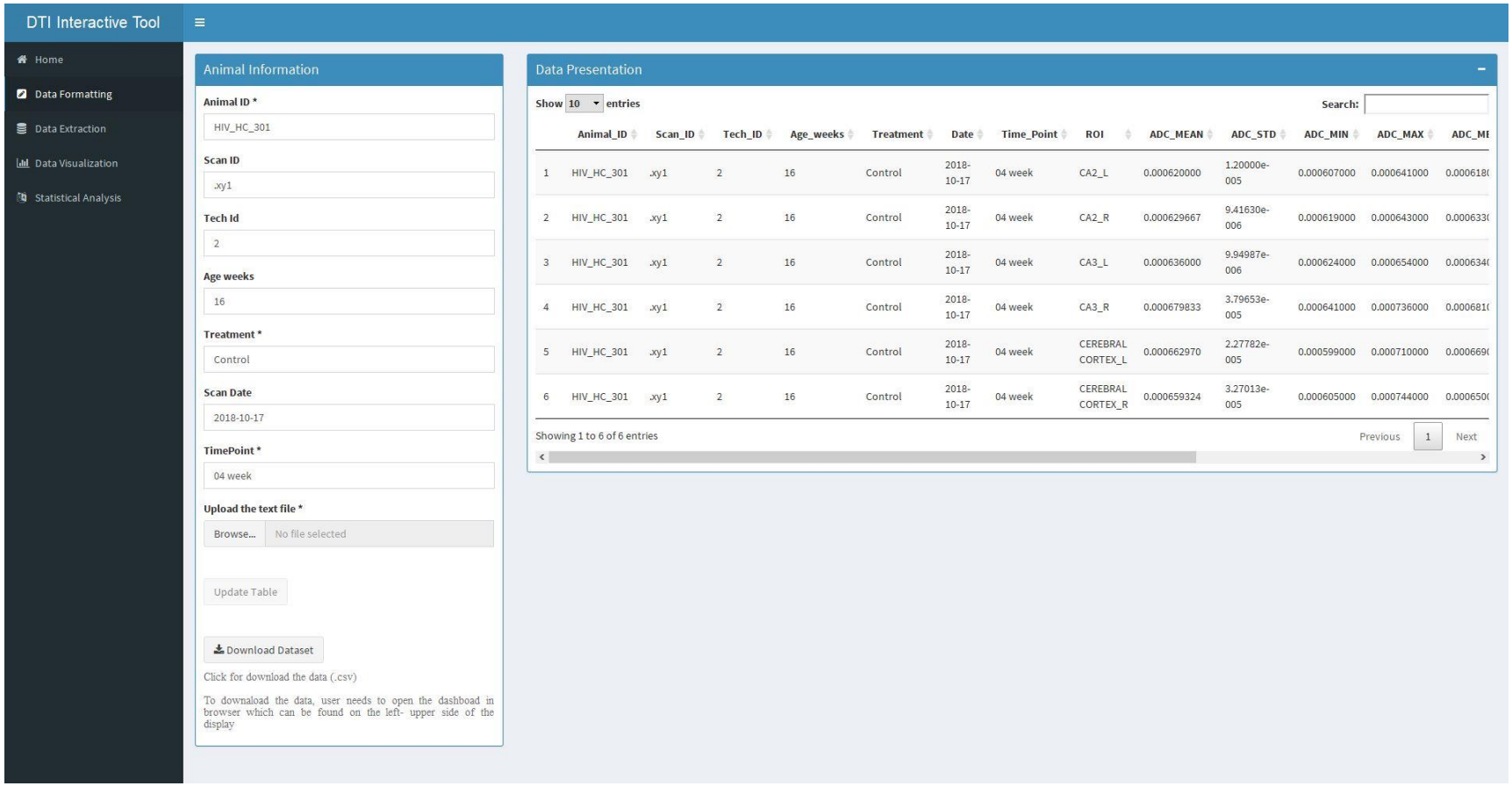


Figure 3.1: Output of the “Data Formatting” tab set. The tab set merged the “HIV_HC_301” animal information and the loaded text file for different ROI (“CA2”, “CA3”, “Cerebral Cortex”) by the user and displayed on “Data Presentation” box.

To extract the data set from the newly-built database, **“Data Extraction”** tab set was developed. “Data Extraction” collects and displays the data in two modes. “Default Extraction” option contains all rows and selective columns from the DTI table in the database whereas “Selective Extraction” option provides the animal name-based dataset which can be used in “Data Visualization” tab set.

Figures 3.2, and 3.3 demonstrate the data extraction procedure by the interactive tool. In Figure 3.2, it shows all rows retrieved from the DTI table with “Animal_id”, “Treatment”, “Time_Point”, “ROI”, “FA_MEAN”, “ADC_MEAN”, and “RA_MEAN” columns. After the extraction, animal names are available for selective extraction. As an example of selective extraction, when an animal name “HIV_HC_302” is selected from the list, then the relevant data rows are shown in the display box (Figure 3.3).

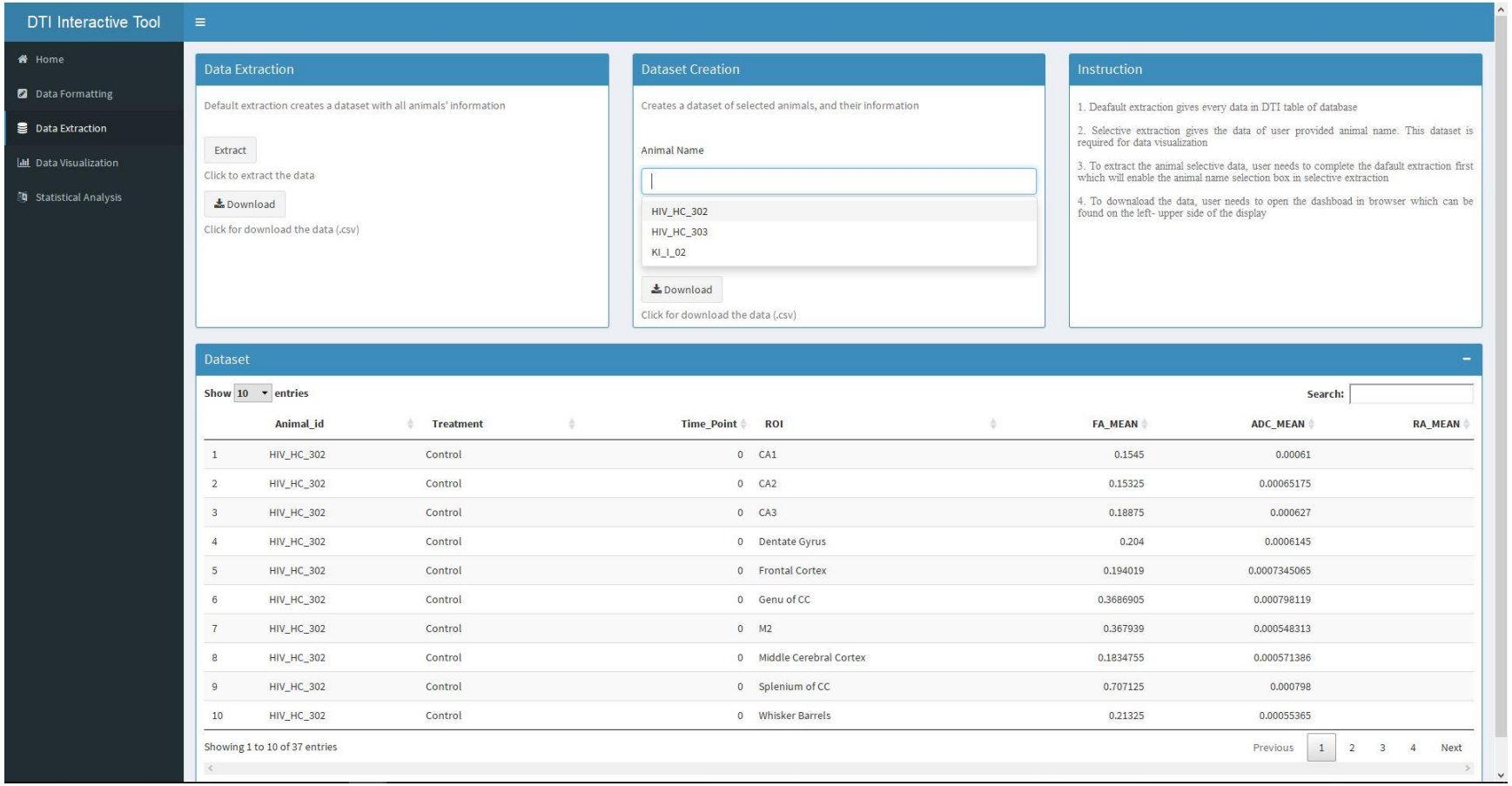


Figure 3.2: Output of “Default Extraction”. This selection also enabled the animal name input box with available animal names for selection.

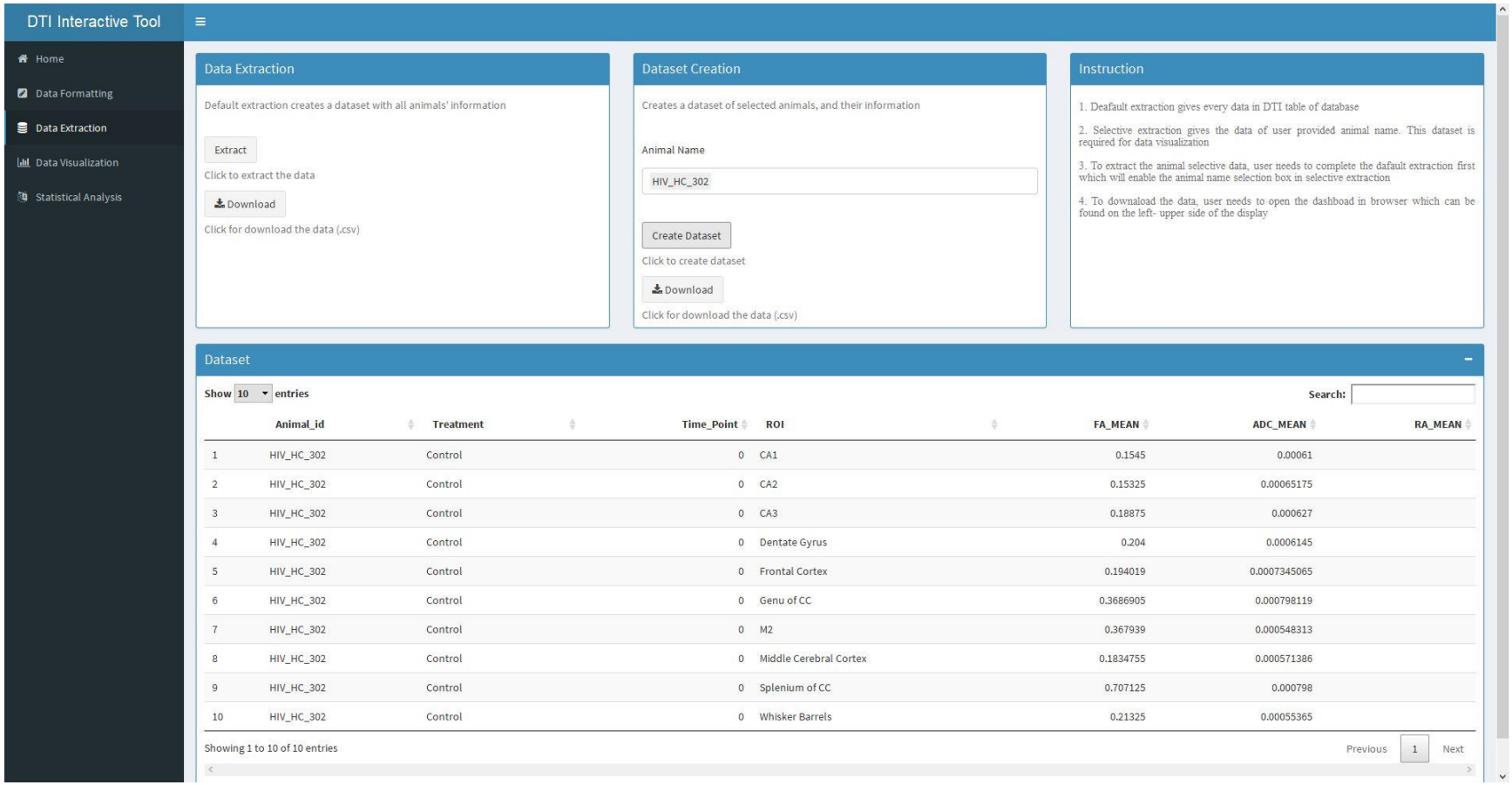


Figure 3.3: Output of “Selective Extraction” where the animal name “HIV_HC_302” is selected.

“Data Visualization” presents box plot, bar plot, and line plot interactively. The user can change any data selection and generate the plots without having to do tedious manual processing. The DTI tool reads the data loaded by the user and prepares the dataset for visual presentation. By using “Data Visualization” tab set, the user can generate plots to qualitatively assess any changes in DTI values over different time points as bar plot and line plot with SEM.

Figure 3.4 shows a dataset in display box which is loaded by the user. That enabled the dynamic selection fields for the user to create plot dataset (figure 3.4 B). In this example, the comparison of FA mean value between “Group A”, “Group B”, and “Group C” at “0 week”, “04 week”, “08 week”, “12 week”, “15 week” time points for “CA1”, and “Cerebellum” regions of the brain is shown. The plots (Figure 3.5) are for visualizing the changes of FA value over those time points.

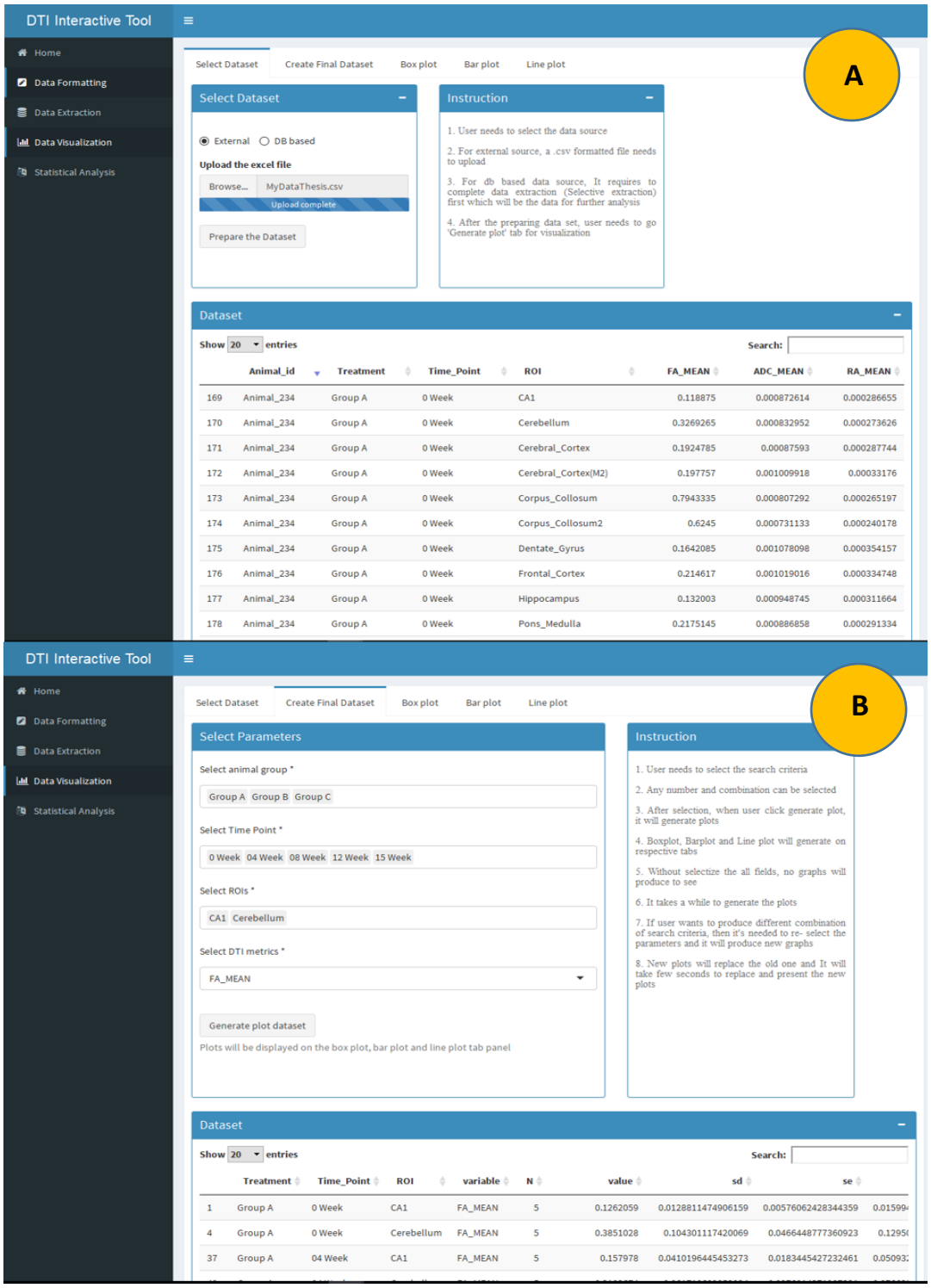


Figure 3.4: “Data Visualization” tab set of the interactive tool. A) Read and display the dataset provided by the user, B) “Create Plot Dataset” subtab set for generating a dataset for displaying, all fields are dynamic. The field’s value changes by changing the dataset.

Figure 3.5: Plots generated by “Data Visualization” tab set. A, B, and C represented the “Box plot”, “Bar plot”, and “Line plot” of “Group A”, “Group B”, and “Group C” at “0 Week”, “04 Week”, “08 Week”, “12 Week”, “15 Week” for CA1 and Cerebellum brain region of the brain.

The last tab set of this interactive tool is **“Statistical Analysis”.** The functionalities include providing dataset summary, and comparing mean between two groups, within the group, between the groups at a specific time point, and all groups with reference to a group. The result from this tab set is a summary of the statistical analysis. In this tab set, the analysis is performed based on the dataset loaded or selected by the user in “Data Visualization” tab set.

Figure 3.6 displays the summary of the dataset. The dataset had a total of 15 animals, and the table represented the number of animals group-wise at different time points. From the summary table, it is found that three DTI metrics were measured from 12 ROIs of the brain. The number and the names are same as the test dataset that is discussed in chapter 2.

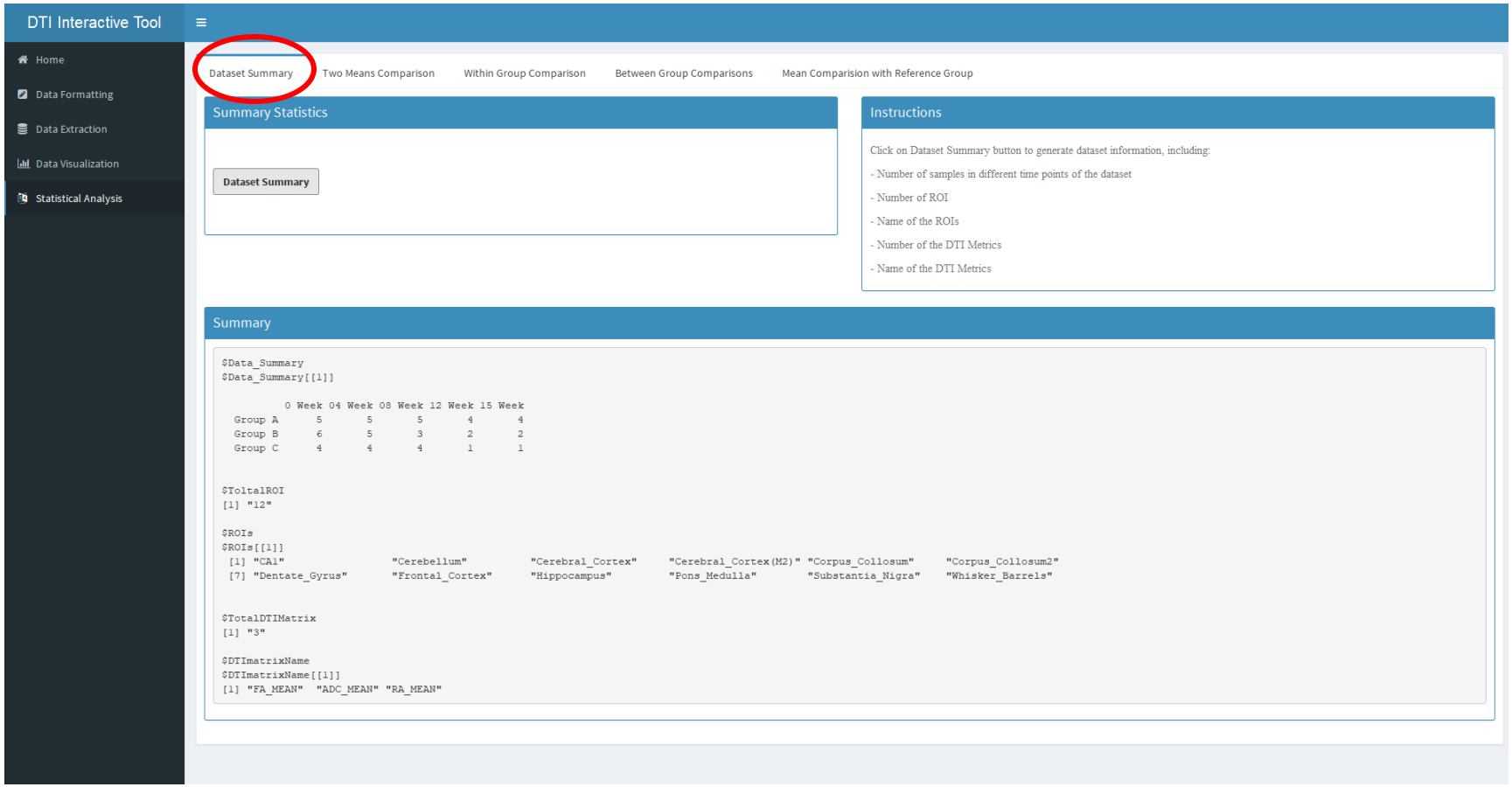


Figure 3.6: Summary of statistical analysis of the dataset including sample size over different time points as group wise with ROI and DTI metrics details.

In Figure 3.7, the summary of two-mean comparison is illustrated. The two-mean comparison of mean FA value for CA1 region of the brain did not show any significant difference between “Group A” and “Group B” at 0 week time point, which was expected because at that time point both groups had the same physiological condition. The P-value from SAS also presented the same result as DTI tool (Figure 3.8).

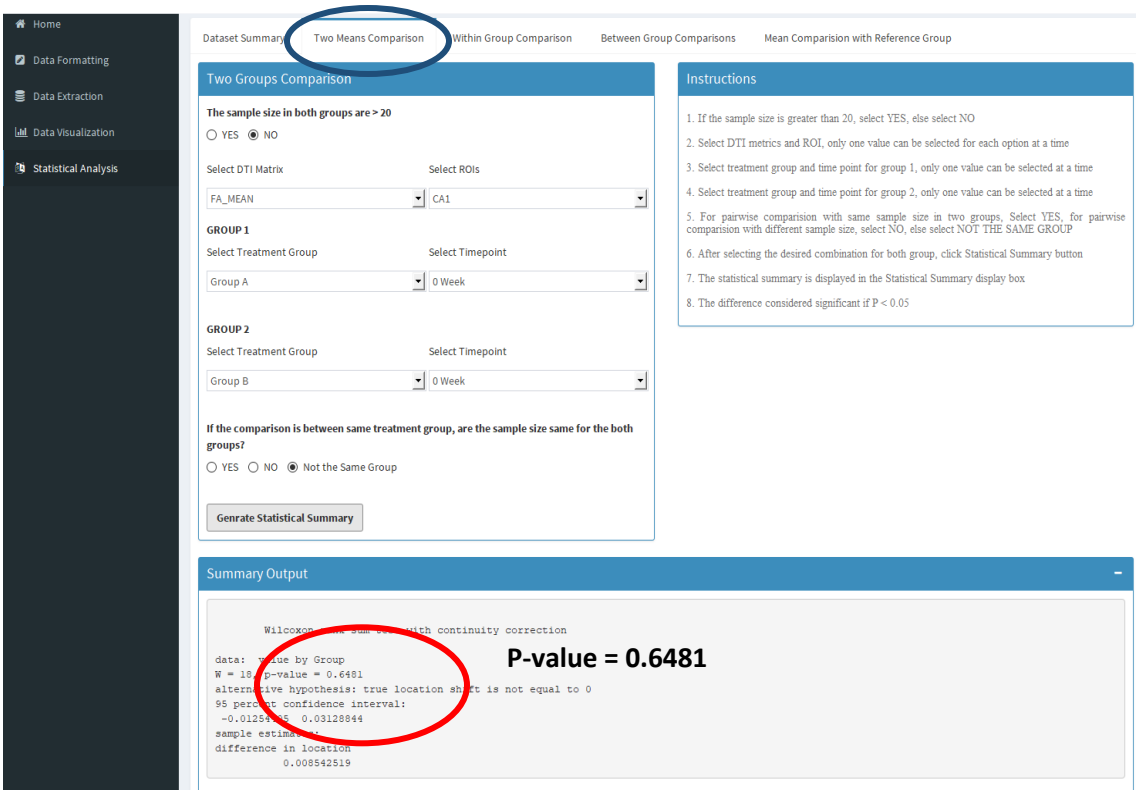


Figure 3.7: Mean comparison output summary from “Statistical Analysis” tab set that shows the result of the comparison of FA between “Group A” and “Group B” at the time point “0 week” for “CA1” region of the brain.

Wilcoxon Scores (Rank Sums) for Variable FA_MEAN Classified by Variable Treatment					
Treatment	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
Group B	6	33.0	36.0	5.477226	5.50
Group A	5	33.0	30.0	5.477226	6.60

Wilcoxon Two-Sample Test	
Statistic	33.0000
Normal Approximation	
Z	0.4564
One-Sided Pr > Z	0.3240
Two-Sided Pr > |Z|	0.6481
t Approximation	
One-Sided Pr > Z	0.3289
Two-Sided Pr > |Z|	0.6578
Z includes a continuity correction of 0.5.	

Figure 3.8: Mean comparison summary from SAS between FA of “Group A” and “Group B” at “0 Week” for “CA1”, where the P-value is 0.6481.

In Figure 3.9, the within “Group B” mean FA value changes over the time for CA1 of the brain region is presented. The result indicates significant differences between “0 week and 08 week”, “0 week and 15 week”, “04 week and 08 week”, “0 week and 15 week”, “08 week and 12 week”, and “12 week” and 15 week”, where the Pr (>|z|) value is below 0.05. The analysis from SAS confirmed these results (Figure 3.10).

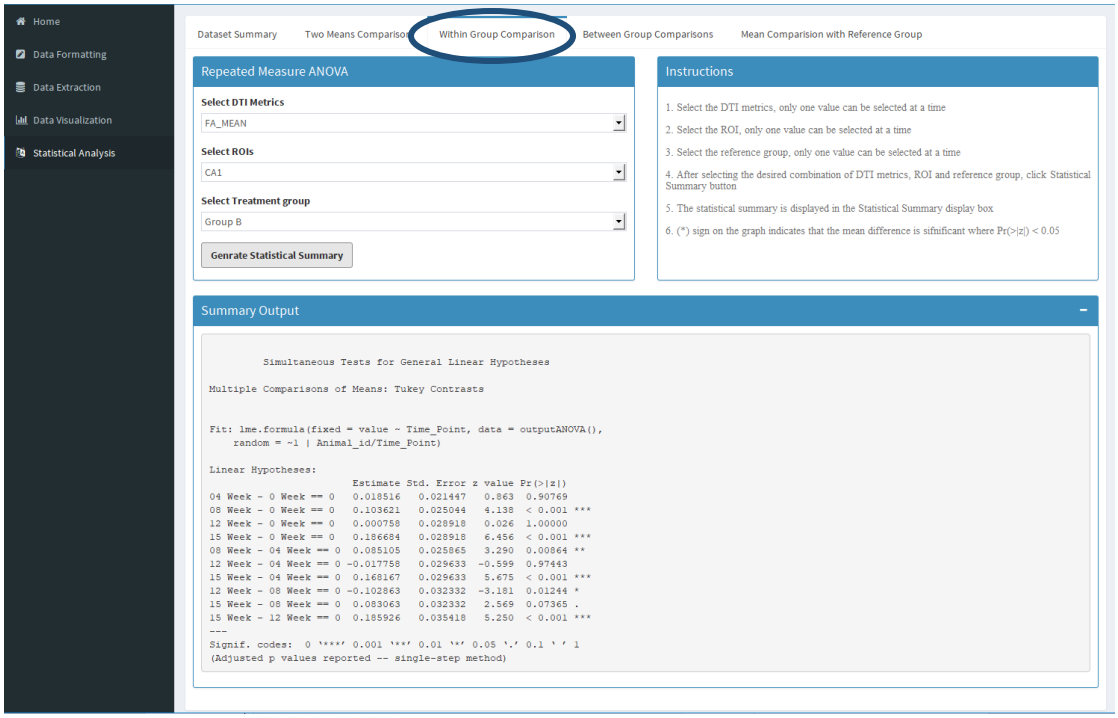
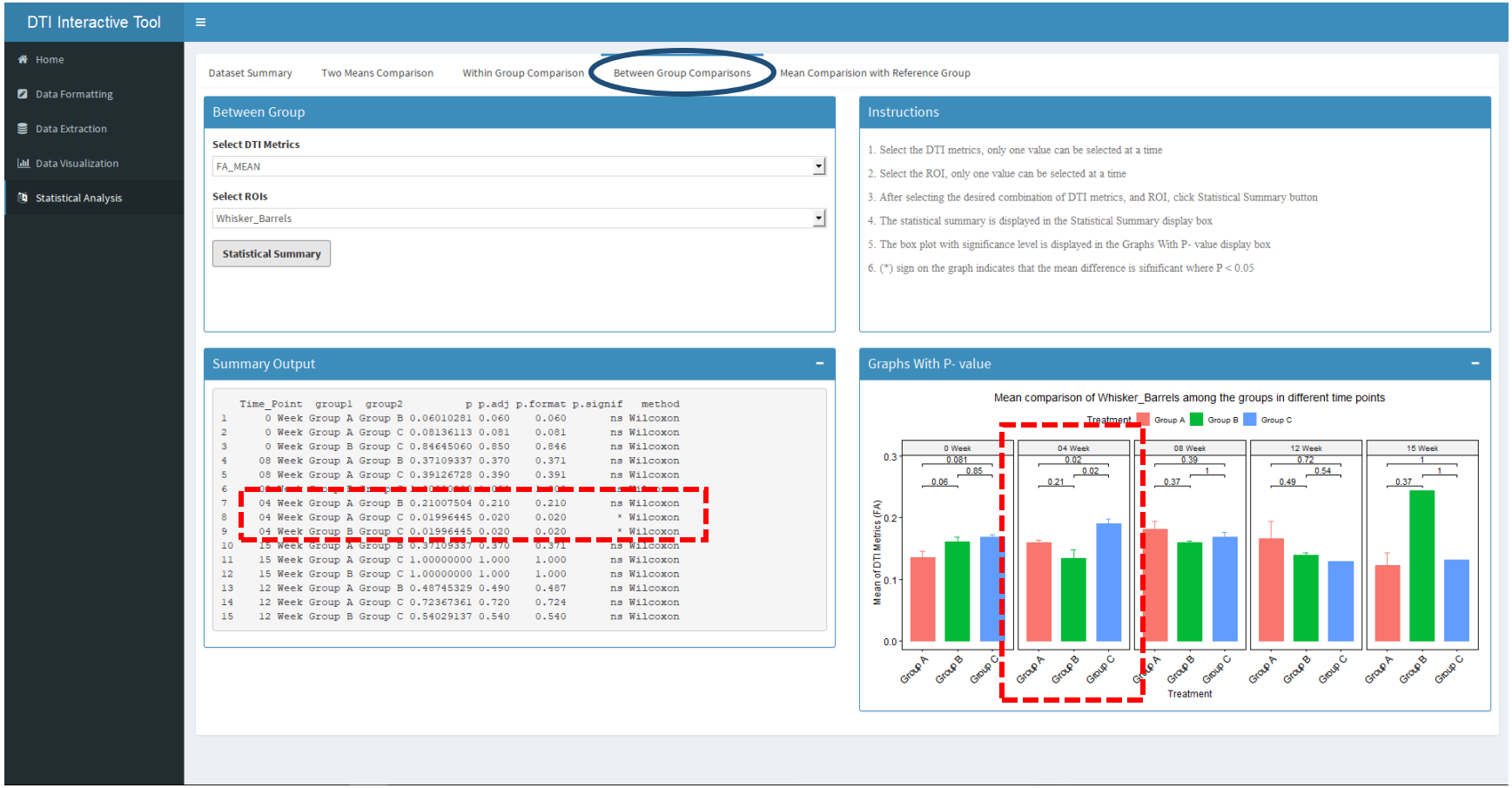


Figure 3.9: Results of within the group comparison of "Group B" for the brain region "CA1" by using repeated measure ANOVA and Tukey's post-hoc test.

Comparisons significant at the 0.05 level are indicated by ***.				
Time_Point Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
15 Week - 08 Week	0.08306	-0.01874	0.18486	
15 Week - 04 Week	0.16817	0.07487	0.26147	***
15 Week - 12 Week	0.18593	0.07441	0.29744	***
15 Week - 0 Week	0.18668	0.09563	0.27774	***
08 Week - 15 Week	-0.08306	-0.18486	0.01874	
08 Week - 04 Week	0.08510	0.00366	0.16655	***
08 Week - 12 Week	0.10286	0.00106	0.20466	***
08 Week - 0 Week	0.10362	0.02477	0.18248	***
04 Week - 15 Week	-0.16817	-0.26147	-0.07487	***
04 Week - 08 Week	-0.08510	-0.16655	-0.00366	***
04 Week - 12 Week	0.01776	-0.07554	0.11106	
04 Week - 0 Week	0.01852	-0.04901	0.08604	
12 Week - 15 Week	-0.18593	-0.29744	-0.07441	***
12 Week - 08 Week	-0.10286	-0.20466	-0.00106	***
12 Week - 04 Week	-0.01776	-0.11106	0.07554	
12 Week - 0 Week	0.00076	-0.09030	0.09181	
0 Week - 15 Week	-0.18668	-0.27774	-0.09563	***
0 Week - 08 Week	-0.10362	-0.18248	-0.02477	***
0 Week - 04 Week	-0.01852	-0.08604	0.04901	
0 Week - 12 Week	-0.00076	-0.09181	0.09030	

Figure 3.10: Statistical summary output of Tukey's post-hoc test by SAS.

Subtab set “Between Group Comparison” compares the mean difference of a DTI metrics among the groups at each time point. Figure 3.11 displays the “FA” means comparison among “Group A”, “Group B”, and “Group C” of “Whisker Barrels” at time points “0 Week”, “04 Week”, “08 Week”, “12 Week”, and “15 Week”. The result indicates that on “04 Week”, there were significant differences between “Group A and Group C” and “Group B and Group C”. The analysis by using SAS for FA mean comparison at time point “04 Week” for “Group B” of “Whisker Barrels” is presented in Figure 3.12, which shows the similar results with DTI tool having exact P-values.



	Time_Point	group1	group2	p	p.adj	p.format	p.signif	method
1	0 Week	Group A	Group B	0.06010281	0.060	0.060	ns	Wilcoxon
2	0 Week	Group A	Group C	0.08136113	0.081	0.081	ns	Wilcoxon
3	0 Week	Group B	Group C	0.84645060	0.850	0.846	ns	Wilcoxon
4	08 Week	Group A	Group B	0.37109337	0.370	0.371	ns	Wilcoxon
5	08 Week	Group A	Group C	0.39126728	0.390	0.391	ns	Wilcoxon
6	[illegible]							
7	04 Week	Group A	Group B	0.21007504	0.210	0.210	ns	Wilcoxon
8	04 Week	Group A	Group C	0.01996445	0.020	0.020	*	Wilcoxon
9	04 Week	Group B	Group C	0.01996445	0.020	0.020	*	Wilcoxon
10	15 Week	Group A	Group B	0.37109337	0.370	0.371	ns	Wilcoxon
11	15 Week	Group A	Group C	1.00000000	1.000	1.000	ns	Wilcoxon
12	15 Week	Group B	Group C	1.00000000	1.000	1.000	ns	Wilcoxon
13	12 Week	Group A	Group B	0.48745329	0.490	0.487	ns	Wilcoxon
14	12 Week	Group A	Group C	0.72367361	0.720	0.724	ns	Wilcoxon
15	12 Week	Group B	Group C	0.54029137	0.540	0.540	ns	Wilcoxon

Figure 3.11: Between the group comparisons of FA mean for Whisker Barrels at multiple time points. At “04 Week”, there are significant difference between “Group A and Group C”, and “Group B and Group C”.

Wilcoxon Scores (Rank Sums) for Variable FA_MEAN Classified by Variable Treatment					
Treatment	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
Group B	5	21.0	27.50	4.787136	4.20
Group A	5	34.0	27.50	4.787136	6.80

Wilcoxon Two-Sample Test	
Statistic	21.0000
Normal Approximation	
Z	-1.2534
One-Sided Pr < Z	0.1050
Two-Sided Pr > |Z|	0.2101
t Approximation	
One-Sided Pr < Z	0.1208
Two-Sided Pr > |Z|	0.2417
Z includes a continuity correction of 0.5.	

Wilcoxon Scores (Rank Sums) for Variable FA_MEAN Classified by Variable Treatment					
Treatment	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
Group C	4	30.0	20.0	4.082483	7.50
Group A	5	15.0	25.0	4.082483	3.00

Wilcoxon Two-Sample Test	
Statistic	30.0000
Normal Approximation	
Z	2.3270
One-Sided Pr > Z	0.0100
Two-Sided Pr > |Z|	0.0200
t Approximation	
One-Sided Pr > Z	0.0242
Two-Sided Pr > |Z|	0.0484
Z includes a continuity correction of 0.5.	

Wilcoxon Scores (Rank Sums) for Variable FA_MEAN Classified by Variable Treatment					
Treatment	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
Group B	5	15.0	25.0	4.082483	3.00
Group C	4	30.0	20.0	4.082483	7.50

Wilcoxon Two-Sample Test	
Statistic	30.0000
Normal Approximation	
Z	2.3270
One-Sided Pr > Z	0.0100
Two-Sided Pr > |Z|	0.0200
t Approximation	
One-Sided Pr > Z	0.0242
Two-Sided Pr > |Z|	0.0484
Z includes a continuity correction of 0.5.	

Figure 3.12: Results from SAS for FA mean comparison between groups at “04 Week” for “Whisker Barrels”.

Subtab set “Mean Comparison with Reference Group” illustrates the comparison of DTI values from all groups at all-time points compared to one individual group, which considered as the reference. Figure 3.13 displays the comparison among the groups where “Group A” at “0 Week” considered as the reference group. The result described that there are significant differences among the reference group and “Group B” and “Group C” at “08 Week” time point and “Group A” at “12 Week” time point for FA mean value of “CA1. These observations matched with SAS results (Figure 3.14).

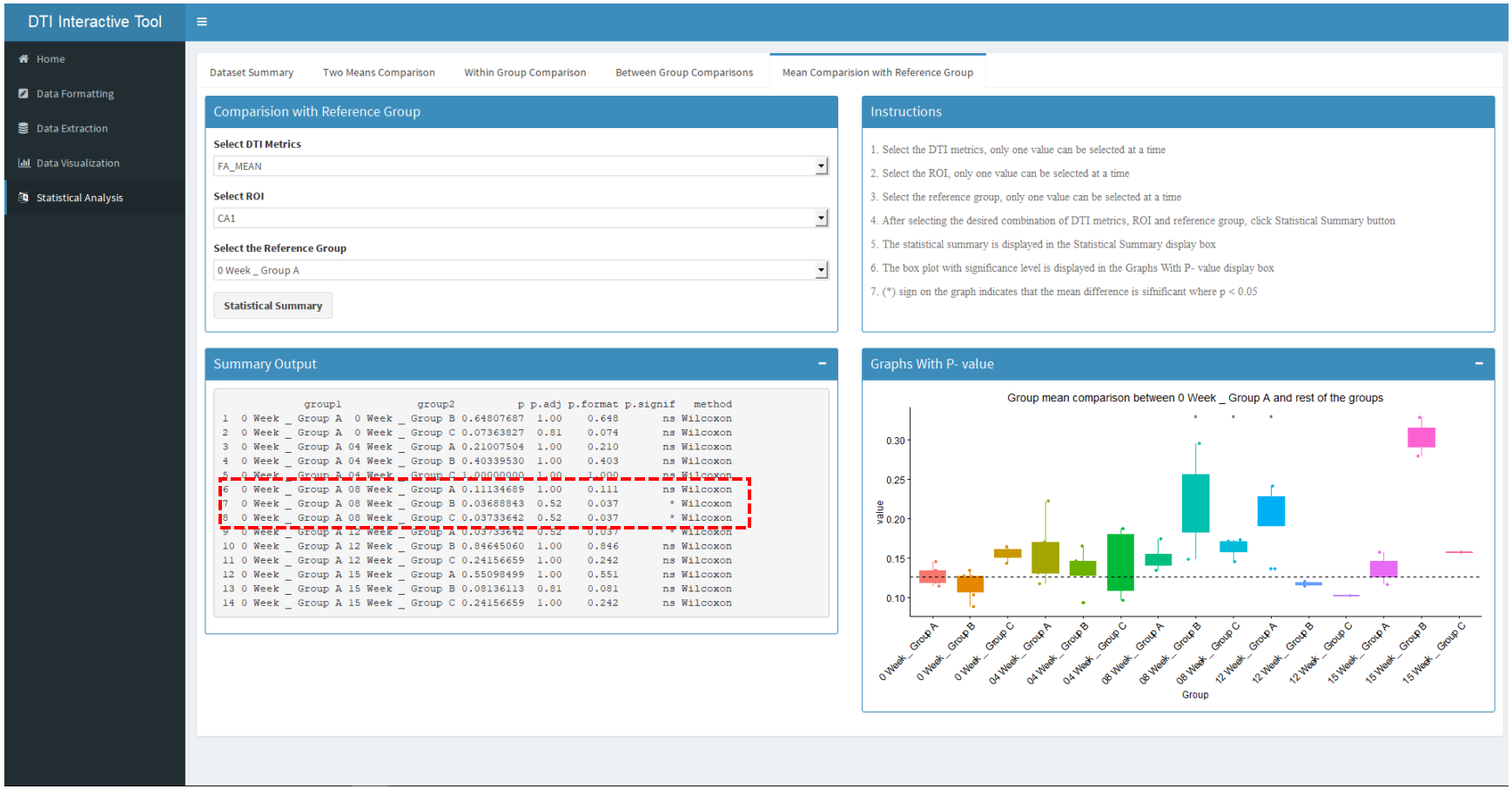


Figure 3.13: Summary output of “Mean Comparison with Reference Group” subtab set where “Group A” at “0 Week” selected as the reference group.

Wilcoxon Scores (Rank Sums) for Variable FA_MEAN Classified by Variable GA0					
GA0	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
0	5	18.0	25.0	4.082483	3.600
1	4	27.0	20.0	4.082483	6.750

Wilcoxon Two-Sample Test	
Statistic	27.0000
Normal Approximation	
Z	1.5922
One-Sided Pr > Z	0.0557
Two-Sided Pr > |Z|	0.1113
t Approximation	
One-Sided Pr > Z	0.0750
Two-Sided Pr > |Z|	0.1500
Z includes a continuity correction of 0.5.	

Wilcoxon Scores (Rank Sums) for Variable FA_MEAN Classified by Variable GA0					
GA0	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
0	5	15.0	22.50	3.354102	3.0
2	3	21.0	13.50	3.354102	7.0

Wilcoxon Two-Sample Test	
Statistic	21.0000
Normal Approximation	
Z	2.0870
One-Sided Pr > Z	0.0184
Two-Sided Pr > |Z|	0.0369
t Approximation	
One-Sided Pr > Z	0.0377
Two-Sided Pr > |Z|	0.0753
Z includes a continuity correction of 0.5.	

Wilcoxon Scores (Rank Sums) for Variable FA_MEAN Classified by Variable GA0					
GA0	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
0	5	16.0	25.0	4.082483	3.200
3	4	29.0	20.0	4.082483	7.250

Wilcoxon Two-Sample Test	
Statistic	29.0000
Normal Approximation	
Z	2.0821
One-Sided Pr > Z	0.0187
Two-Sided Pr > |Z|	0.0373
t Approximation	
One-Sided Pr > Z	0.0354
Two-Sided Pr > |Z|	0.0709
Z includes a continuity correction of 0.5.	

Figure 3.14: Statistical summary of the mean comparison between “Group A” at “0 Week” and “Group A”, “Group B” and “Group C” at “08 Week” where GA0 value 0 denotes “0 Week_Group A”, 1 denotes “08 Week_Group A”, 2 indicates “08 Week_Group B” and 3 is for “0 Week_Group A”.

In this chapter, the newly developed interactive analytical tool was tested for the accuracy and robustness by using simulate test dataset. Results suggested that the tool is reliable to analyze DTI-related measures in preclinical research setup.

CHAPTER 4: DISCUSSION

In the twenty-first century, biology research has been transformed into a data-rich field because of the mammoth amount of data produced from different sources including human, biological substances, and cells and living organisms, using advanced and high-throughput experimental methods. These biological data are collected in broad and diverse formats and consisted of sequences, graphs, metabolomics pathways, protein structures, images, research studies and techniques, experimental output, scientific literature, etc. (Wooley & Lin, 2005). To collect, manage and access the biological data with accuracy and consistency, the concept of biological databases has evolved. Many such biological databases were developed to serve specific purposes (Brynne et al., 2013). Preclinical research utilizes the biological, biochemical, structural, functional and behavioral information to gain insights into pathobiology and treatment opportunities for human disease. For this purpose, multiple modalities are used to acquire multidimensional data. In order to interpret all the metrics from these data together, a database and interactive tools for data analytics are essential. Tools that can automatically process different formats of data, provide analytics, and bring the measures into a common coordinate system by aligning them to standard atlas will greatly help preclinical research community.

In this thesis, we presented a model of the database and interactive analytical tool for DTI data visualization and statistical analysis in a preclinical research setting. The present work was performed to achieve three objectives. The first objective was to develop a standard data storing method to replace the current manual storing procedure to extract any information efficiently. The second objective was to reduce the human-related errors while handling the data. The third one was to improve the data visualization process and decrease the data analysis time, and the reduced dependency on other software

packages like SAS, SPSS, and Microsoft Office for generating analytical graphs, plots and statistical summary by automating the processing steps. Our DTI interactive tool and database showed significant improvements in data formatting and analyzing effort. Previously, it required more than an hour to format one set of animal data and three days to prepare the data for analyses and produce plots and statistical summary by following the conventional method, whilst using the present DTI analytical tool it took less than **ten minutes** to produce the same results interactively without using multiple analytical software. That greatly helps principal investigators to quickly interpret current experimental results and plan efficiently for future studies.

MR-DTI is a non-invasive and *in vivo* technique that often used to probe the microstructural integrity of brain by in vivo measuring the water diffusivity. Thus, it is also considered as a powerful tool in neurodegenerative research in small animals. To store and present the quantitative DTI data, we have designed a new entity-relational database and a user-friendly, interactive DTI analytical tool. The DTI tool formatted the text data for inserting into a relational database, retrieved the required dataset from the database without going through manual intervention, and presented the graphs, plots and statistical summary interactively which reduced the formatting and analyzing time significantly. Matched analysis results from our Tool with SAS results on a test dataset confirmed the correct implementation of the analytical tool. Microsoft SQL Server 2016 Express Edition was used as the DBMS, and an open source programming language, R, for developing the interactive DTI analytical tool. Due to the stand-alone nature of the developed DTI tool, users can also access this tool through the R server, and analyze their own data without connecting to the database.

In the present thesis work, DTI interactive tool was developed as a prototype that extracted data from the DTI table of the existing database. In future, other MR quantitative

measures such as biochemical profiles from spectroscopy, T1 and T2 relaxation times that are sensitive to biological changes along with biological information such as disease information, histological data, and data related to viral load will be integrated into the interactive tool and can explore applying supervised and unsupervised machine learning techniques to predict the outcome of the experiments. Due to lack of automated tools, researchers usually perform analysis of quantitative measures only in selected ROIs and thus not fully utilizing the data acquired. However, by registering all MRI data to a common average MRI atlas (Sajja et al. 2016), it is possible to perform automatic whole brain analysis and increase the strength of the analysis. In future, we will be integrating population average MRI atlas into our Analytical Tool.

REFERENCES

Aban, I. B., & George, B. (2015). Statistical considerations for preclinical studies. Experimental Neurology, 270, 82-87. doi:10.1016/j.expneurol.2015.02.024

Alexander, A. L., Lee, J. E., Lazar, M., & Field, A. S. (2007). Diffusion tensor imaging of the brain. Neurotherapeutics,4(3), 316-329. doi:10.1016/j.nurt.2007.05.011

Badgeley, M. A., Shameer, K., Glicksberg, B. S., Tomlinson, M. S., Levin, M. A., Mccormick, P. J., . . . Dudley, J. T. (2016). EHDViz: Clinical dashboard development using open-source technologies. BMJ Open,6(3). doi:10.1136/bmjopen-2015-010579

Basser, P. J., & Pierpaoli, C. (1996). Microstructural and Physiological Features of Tissues Elucidated by Quantitative-Diffusion-Tensor MRI. Journal of Magnetic Resonance, Series B,111(3), 209-219. doi:10.1006/jmrb.1996.0086

Basser, P., Mattiello, J., & Lebihan, D. (1994). Estimation of the Effective Self-Diffusion Tensor from the NMR Spin Echo. Journal of Magnetic Resonance, Series B,103(3), 247-254. doi:10.1006/jmrb.1994.1037

Bihan, D. L., Mangin, J., Poupon, C., Clark, C. A., Pappata, S., Molko, N., & Chabriat, H. (2001). Diffusion tensor imaging: Concepts and applications. Journal of Magnetic Resonance Imaging,13(4), 534-546. doi:10.1002/jmri.1076

Boska, M. D., Dash, P. K., Knibbe, J., Epstein, A. A., Akhter, S. P., Fields, N., . . . Gorantla, S. (2014). Associations between brain microstructures, metabolites, and cognitive deficits during chronic HIV-1 infection of humanized mice. Molecular Neurodegeneration,9(1), 58. doi:10.1186/1750-1326-9-58

Corder, G. W., & Foreman, D. I. (2009). NONPARAMETRIC STATISTICS: AN INTRODUCTION. In Nonparametric statistics for non-statisticians: A step-by-step approach (p. 2). Hoboken, NJ: Wiley.

Coronel, C., Morris, S., & Rob, P. (2011). Chapter 2: Data Models. In Database systems: Design, implementation, and management. Australia: Course Technology Cengage Learning.

Emsell, L., Hecke, W. V., & Tournier, J. (2016). Introduction to Diffusion Tensor Imaging. Diffusion Tensor Imaging,7-19. doi:10.1007/978-1-4939-3118-7_2

Gorantla, S., Makarov, E., Finke-Dwyer, J., Gebhart, C. L., Domm, W., Dewhurst, S., . . . Poluektova, L. Y. (2010, June 15). CD8 Cell Depletion Accelerates HIV-1 Immunopathology in Humanized Mice. Retrieved from http://www.jimmunol.org/content/184/12/7082

Heuvel, M. P., Mandl, R. C., Kahn, R. S., & Pol, H. E. (2009). Functionally linked resting-state networks reflect the underlying structural connectivity architecture of the human brain. Human Brain Mapping,30(10), 3127-3141. doi:10.1002/hbm.

Hoffer, J. A., Ramesh, V., & Topi, H. (2011). Chapter 1: The Database Environment and Development Process. In Modern database management (pp. 5-6). Upper Saddle River, NJ: Prentice Hall.

Huisman, T. (2010). Diffusion-weighted and diffusion tensor imaging of the brain, made easy. Cancer Imaging,10(1A). doi:10.1102/1470-7330.2010.9023

Kroenke, D., & Auer, D. J. (2012). Chapter 1: Introduction. In Database processing: Fundamentals, design, and implementation (p. 8). Boston: Pearson.

Lazar, M. (2010). Mapping brain anatomical connectivity using white matter tractography. NMR in Biomedicine,23(7), 821-835. doi:10.1002/nbm.1579

Martin, Greg S. “The Essential Nature of Healthcare Databases in Critical Care Medicine.” Critical Care, vol. 12, no. 5, 2008, p. 176., doi:10.1186/cc6993.

Mori, S., & Zhang, J. (2006). Principles of Diffusion Tensor Imaging and Its Applications to Basic Neuroscience Research. Neuron,51(5), 527-539. doi:10.1016/j.neuron.2006.08.012

Mori, S., & Zhang, J. (2006). Principles of Diffusion Tensor Imaging and Its Applications to Basic Neuroscience Research. Neuron,51(5), 527-539. doi:10.1016/j.neuron.2006.08.012

Mukherjee, P., Berman, J., Chung, S., Hess, C., & Henry, R. (2008). Diffusion Tensor MR Imaging and Fiber Tractography: Theoretic Underpinnings. American Journal of Neuroradiology,29(4), 632-641. doi:10.3174/ajnr.a1051

O'Donnell, L. J., & Westin, C. (2011). An Introduction to Diffusion Tensor Image Analysis. Neurosurgery Clinics of North America,22(2), 185-196. doi:10.1016/j.nec.2010.12.004

Pierpaoli, C., Jezzard, P., Basser, P. J., Barnett, A., & Chiro, G. D. (1996). Diffusion tensor MR imaging of the human brain. Radiology,201(3), 637-648. doi:10.1148/radiology.201.3.8939209

Sajja, B. R., Bade, A. N., Zhou, B., Uberti, M. G., Gorantla, S., Gendelman, H. E., ... Liu, Y. (Accepted/In press). Generation and Disease Model Relevance of a Manganese Enhanced Magnetic Resonance Imaging-Based NOD/scid-IL-2Rγ_c ^{null} Mouse Brain Atlas. *Journal of NeuroImmune Pharmacology*. https://doi.org/10.1007/s11481-015-9635-8

Salahuddin, P., & Khan, A. U. (2008). Proteolytic Enzymes Database. Journal of Proteomics & Bioinformatics,01(02), 109-111. doi:10.4172/jpb.1000017

Sirianni, A. The Importance of Data and Information in Business - Blog - DCODE GROUP - DCODE GROUP. Retrieved from https://www.dcodegroup.com/blog/the-importance-of-data-and-information-in-business

SQL. (2018, October 24). Retrieved from https://en.wikipedia.org/wiki/SQL

Williams, E., Moore, J., Li, S. W., Rustici, G., Tarkowska, A., Chessel, A., . . . Swedlow, J. R. (2017). Image Data Resource: A bioimage data integration and publication platform. Nature Methods,14(8), 775-781. doi:10.1038/nmeth.4326

Wooley, J. C., & Lin, H. (2005). Catalyzing inquiry at the interface of computing and biology. National Academies Press.

APPENDIX A

ANIMAL						
Column Name	Data type	Max Length	precision	scale	is_nullable	Primary Key
Animal_Id	varchar	100	0	0	0	1
Animal_Type	varchar	100	0	0	1	0
Strain	varchar	50	0	0	1	0
Gender	varchar	10	0	0	1	0
DOB	date	3	10	0	1	0
Irr_date	date	3	10	0	1	0
Inj_date	date	3	10	0	1	0
End_date	date	3	10	0	1	0
Scan_date	date	3	10	0	1	0
Cage_Id	int	4	10	0	1	0
Dis_id	int	4	10	0	1	0
PI_Id	int	4	10	0	1	0
Stu_Name	varchar	100	0	0	1	0
Comments	varchar	100	0	0	1	0
Dest_Id	int	4	10	0	1	0
Scan_Start_Time	varchar	50	0	0	1	0
Scan_End_Time	varchar	50	0	0	1	0
Treatment_Group	varchar	100	0	0	1	0
Time_Point	varchar	50	0	0	1	0
IACUC_Number	int	4	10	0	1	0

BLEED						
Column Name	Data type	Max Length	precision	scale	is_nullable	Primary Key

Id	int	4	10	0	0	1
Animal_id	varchar	100	0	0	1	0
Bleed_date	date	3	10	0	1	0
Tech_id	int	4	10	0	1	0
cd45	numeric	9	18	0	1	0
cd3	numeric	9	18	0	1	0
cd19	numeric	9	18	0	1	0
cd4	numeric	9	18	0	1	0
cd8	int	4	10	0	1	0
cd14	numeric	9	18	0	1	0
cd45ra	numeric	9	18	0	1	0
cd4ra	numeric	9	18	0	1	0
cd8ra	numeric	9	18	0	1	0
viral_load	numeric	9	18	0	1	0
Comments	varchar	200	0	0	1	0

COST CENTER						
Column Name	Data type	Max Length	precision	scale	is_nullable	Primary Key
Id	int	4	10	0	0	1
CC_Number	int	4	10	0	1	0
Stu_Name	varchar	100	0	0	1	0

DESTINATION						
Column Name	Data type	Max Length	precision	scale	is_nullable	Primary Key
Dest_ID	int	4	10	0	0	1
Description	varchar	100	0	0	1	0

DISEASE						
Column Name	Data type	Max Length	precision	scale	is_nullable	Primary Key
Dis_ID	int	4	10	0	0	1
Dis_Name	varchar	100	0	0	1	0
Dis_Prod_Date	date	3	10	0	1	0
Dis_Conc	varchar	100	0	0	1	0
Dis_Date	date	3	10	0	1	0
Dis_Source	varchar	100	0	0	1	0
Comments	varchar	100	0	0	1	0

DTI						
Column Name	Data type	Max Length	precision	scale	is_nullable	Primary Key
Id	int	4	10	0	0	1
Animal_id	varchar	100	0	0	1	0
Scan_Id	varchar	50	0	0	1	0
Tech_Id	int	4	10	0	1	0
Age_Weeks	int	4	10	0	1	0
Treatment	varchar	50	0	0	1	0
Date	date	3	10	0	1	0
Time_Point	int	4	10	0	1	0
ROI_Location	varchar	100	0	0	1	0
ADC_MEAN	float	8	53	0	1	0
ADC_SD	float	8	53	0	1	0
ADC_MIN	float	8	53	0	1	0
ADC_MAX	float	8	53	0	1	0
ADC_MED	float	8	53	0	1	0

FA_MEAN	float	8	53	0	1	0
FA_SD	float	8	53	0	1	0
FA_MIN	float	8	53	0	1	0
FA_MAX	float	8	53	0	1	0
FA_MED	float	8	53	0	1	0
Dax_MEAN	float	8	53	0	1	0
Dax_SD	float	8	53	0	1	0
Drad_MEAN	float	8	53	0	1	0
Drad_SD	float	8	53	0	1	0
RA_MEAN	float	8	53	0	1	0
RA_SD	float	8	53	0	1	0
CL_mean	float	8	53	0	1	0
CL_SD	float	8	53	0	1	0
CP_MEAN	float	8	53	0	1	0
CP_SD	float	8	53	0	1	0
CS_MEAN	float	8	53	0	1	0
CS_SD	float	8	53	0	1	0
File_Location	varchar	1000	0	0	1	0

HISTOLOGY						
Column Name	Data type	Max Length	precision	scale	is_nullable	Primary Key
Id	int	4	10	0	0	1
Region	varchar	50	0	0	1	0
Stain_type	varchar	100	0	0	1	0
Stain_quantity	varchar	50	0	0	1	0
Animal_id	varchar	100	0	0	1	0

IACUC						
Column Name	Data type	Max Length	precision	scale	is_nullable	Primary Key
Iacuc_Number	int	4	10	0	0	1
Stu_Name	varchar	100	0	0	1	0

LC MODEL						
Column Name	Data type	Max Length	precision	scale	is_nullable	Primary Key
Id	int	4	10	0	0	1
Scan_ID	varchar	50	0	0	1	0
Treatment	varchar	50	0	0	1	0
Time_Point	int	4	10	0	1	0
Time_Point_U	varchar	100	0	0	1	0
Area_Name	varchar	100	0	0	1	0
Area_Dir	float	8	53	0	1	0
ALA_A	float	8	53	0	1	0
ASP_A	float	8	53	0	1	0
CR_A	float	8	53	0	1	0
GABA_A	float	8	53	0	1	0
GLC_A	float	8	53	0	1	0
GLN_A	float	8	53	0	1	0
GLU_A	float	8	53	0	1	0
GLY_A	float	8	53	0	1	0
GPC_A	float	8	53	0	1	0
LAC_A	float	8	53	0	1	0
mI_A	float	8	53	0	1	0
NAA_A	float	8	53	0	1	0

PC_A	float	8	53	0	1	0
TAU_A	float	8	53	0	1	0
tCHO_A	float	8	53	0	1	0
ALA_SD	float	8	53	0	1	0
ASP_SD	float	8	53	0	1	0
CR_SD	float	8	53	0	1	0
GABA_SD	float	8	53	0	1	0
GLC_SD	float	8	53	0	1	0
GLN_SD	float	8	53	0	1	0
GLU_SD	float	8	53	0	1	0
GLY_SD	float	8	53	0	1	0
GPC_SD	float	8	53	0	1	0
LAC_SD	float	8	53	0	1	0
mI_SD	float	8	53	0	1	0
NAA_SD	float	8	53	0	1	0
PC_SD	float	8	53	0	1	0
TAU_SD	float	8	53	0	1	0
tCHO_SD	float	8	53	0	1	0
ALA_A_PTS	float	8	53	0	1	0
ASP_A_PTS	float	8	53	0	1	0
CR_A_PTS	float	8	53	0	1	0
GABA_A_PTS	float	8	53	0	1	0
GLC_A_PTS	float	8	53	0	1	0
GLN_A_PTS	float	8	53	0	1	0
GLU_A_PTS	float	8	53	0	1	0
GLY_A_PTS	float	8	53	0	1	0

GPC_A_PTS	float	8	53	0	1	0
LAC_A_PTS	float	8	53	0	1	0
mI_A_PTS	float	8	53	0	1	0
NAA_A_PTS	float	8	53	0	1	0
PC_A_PTS	float	8	53	0	1	0
TAU_A_PTS	float	8	53	0	1	0
tCHO_A_PTS	float	8	53	0	1	0
ALA_SD_PTS	float	8	53	0	1	0
ASP_SD_PTS	float	8	53	0	1	0
CR_SD_PTS	float	8	53	0	1	0
GABA_SD_PTS	float	8	53	0	1	0
GLC_SD_PTS	float	8	53	0	1	0
GLN_SD_PTS	float	8	53	0	1	0
GLU_SD_PTS	float	8	53	0	1	0
GLY_SD_PTS	float	8	53	0	1	0
GPC_SD_PTS	float	8	53	0	1	0
LAC_SD_PTS	float	8	53	0	1	0
mI_SD_PTS	float	8	53	0	1	0
NAA_SD_PTS	float	8	53	0	1	0
PC_SD_PTS	float	8	53	0	1	0
TAU_SD_PTS	float	8	53	0	1	0
tCHO_SD_PTS	float	8	53	0	1	0
Animal_id	varchar	100	0	0	1	0

MODEL						
Column Name	Data type	Max Length	precision	scale	is_nullable	Primary Key

Id	int	4	10	0	0	1
Dis_id	int	4	10	0	1	0
Animal_id	varchar	100	0	0	1	0

MORPHOMETRY						
Column Name	Data type	Max Length	precision	scale	is_nullable	Primary Key
Id	int	4	10	0	0	1
Animal_id	varchar	100	0	0	1	0
Region	varchar	100	0	0	1	0
Appl	varchar	100	0	0	1	0
Measurement	float	8	53	0	1	0
Comments	varchar	400	0	0	1	0

NANOMATERIALS						
Column Name	Data type	Max Length	precision	scale	is_nullable	Primary Key
Id	int	4	10	0	0	1
Batch_id	int	4	10	0	1	0
Animal_id	varchar	100	0	0	1	0
Tech_id	int	4	10	0	1	0
Comments	varchar	400	0	0	1	0

PI						
Column Name	Data type	Max Length	precision	scale	is_nullable	Primary Key
PI_ID	int	4	10	0	0	1
First_Name	varchar	100	0	0	1	0
Last_Name	varchar	100	0	0	1	0

Email	varchar	100	0	0	1	0
Contact_Number	int	4	10	0	1	0

QUEST						
Column Name	Data type	Max Length	precision	scale	is_nullable	Primary Key
Id	int	4	10	0	0	1
Scan_ID	varchar	50	0	0	1	0
Treatment	varchar	50	0	0	1	0
Time_Point	int	4	10	0	1	0
Time_Point_U	varchar	100	0	0	1	0
Area_Name	varchar	100	0	0	1	0
Area_Dir	float	8	53	0	1	0
ALA_A	float	8	53	0	1	0
ASP_A	float	8	53	0	1	0
CR_A	float	8	53	0	1	0
GABA_A	float	8	53	0	1	0
GLC_A	float	8	53	0	1	0
GLN_A	float	8	53	0	1	0
GLU_A	float	8	53	0	1	0
GLY_A	float	8	53	0	1	0
GPC_A	float	8	53	0	1	0
LAC_A	float	8	53	0	1	0
mI_A	float	8	53	0	1	0
NAA_A	float	8	53	0	1	0
PC_A	float	8	53	0	1	0

TAU_A	float	8	53	0	1	0
tCHO_A	float	8	53	0	1	0
ALA_SD	float	8	53	0	1	0
ASP_SD	float	8	53	0	1	0
CR_SD	float	8	53	0	1	0
GABA_SD	float	8	53	0	1	0
GLC_SD	float	8	53	0	1	0
GLN_SD	float	8	53	0	1	0
GLU_SD	float	8	53	0	1	0
GLY_SD	float	8	53	0	1	0
GPC_SD	float	8	53	0	1	0
LAC_SD	float	8	53	0	1	0
mI_SD	float	8	53	0	1	0
NAA_SD	float	8	53	0	1	0
PC_SD	float	8	53	0	1	0
TAU_SD	float	8	53	0	1	0
tCHO_SD	float	8	53	0	1	0
ALA_A_PTS	float	8	53	0	1	0
ASP_A_PTS	float	8	53	0	1	0
CR_A_PTS	float	8	53	0	1	0
GABA_A_PTS	float	8	53	0	1	0
GLC_A_PTS	float	8	53	0	1	0
GLN_A_PTS	float	8	53	0	1	0
GLU_A_PTS	float	8	53	0	1	0
GLY_A_PTS	float	8	53	0	1	0
GPC_A_PTS	float	8	53	0	1	0

LAC_A_PTS	float	8	53	0	1	0
mI_A_PTS	float	8	53	0	1	0
NAA_A_PTS	float	8	53	0	1	0
PC_A_PTS	float	8	53	0	1	0
TAU_A_PTS	float	8	53	0	1	0
tCHO_A_PTS	float	8	53	0	1	0
ALA_SD_PTS	float	8	53	0	1	0
ASP_SD_PTS	float	8	53	0	1	0
CR_SD_PTS	float	8	53	0	1	0
GABA_SD_PTS	float	8	53	0	1	0
GLC_SD_PTS	float	8	53	0	1	0
GLN_SD_PTS	float	8	53	0	1	0
GLU_SD_PTS	float	8	53	0	1	0
GLY_SD_PTS	float	8	53	0	1	0
GPC_SD_PTS	float	8	53	0	1	0
LAC_SD_PTS	float	8	53	0	1	0
mI_SD_PTS	float	8	53	0	1	0
NAA_SD_PTS	float	8	53	0	1	0
PC_SD_PTS	float	8	53	0	1	0
TAU_SD_PTS	float	8	53	0	1	0
tCHO_SD_PTS	float	8	53	0	1	0
Animal_id	varchar	100	0	0	1	0

RESEARCH						
Column Name	Data type	Max Length	precision	scale	is_nullable	Primary Key
Id	int	4	10	0	0	1

PI_Id	int	4	10	0	1	0
Iacuc_Number	int	4	10	0	1	0

SCAN TYPE						
Column Name	Data type	Max Length	precision	scale	is_nullable	Primary Key
Id	int	4	10	0	0	1
Scan_Type	varchar	100	0	0	1	0
Animal_Id	varchar	100	0	0	1	0

SPECTRA						
Column Name	Data type	Max Length	precision	scale	is_nullable	Primary Key
Spectra_Id	int	4	10	0	0	1
Stu_Name	varchar	100	0	0	1	0
Number_of_Subjects	int	4	10	0	1	0
Quantification_Method	varchar	50	0	0	1	0
Water_Scaling	varchar	100	0	0	1	0
Eddy_Current_Correction	varchar	100	0	0	1	0
Correction_Factor	float	8	53	0	1	0

STUDY						
Column Name	Data type	Max Length	precision	scale	is_nullable	Primary Key
Stu_Name	varchar	100	0	0	0	1
PI_ID	int	4	10	0	1	0

TECHNICIAN

Column Name	Data type	Max Length	precision	scale	is_nullable	Primary Key
Tech_ID	int	4	10	0	0	1
First_Name	varchar	100	0	0	1	0
Last_Name	varchar	100	0	0	1	0
Email_Address	varchar	50	0	0	1	0
Comments	varchar	400	0	0	1	0